

# Hybrid machine learning data assimilation for marine biogeochemistry

Ieuan Higgs<sup>1,2</sup>, Ross Bannister<sup>1,2</sup>, Jozef Skákala<sup>2,3</sup>, Alberto Carrassi<sup>1,4</sup>, and Stefano Ciavatta<sup>5</sup>

<sup>1</sup>Department of Meteorology, University of Reading, UK

<sup>2</sup>National Centre for Earth Observation, UK

<sup>3</sup>Plymouth Marine Laboratory, UK

<sup>4</sup>Department of Physics and Astronomy “Augusto Righi”, University of Bologna, IT

<sup>5</sup>Mercator Ocean International, FR

November 13, 2025

Corresponding Author: Ieuan Higgs (i.c.higgs@reading.ac.uk)

## Abstract

Marine biogeochemistry models are critical for forecasting, as well as estimating ecosystem responses to climate change and human activities. Data assimilation (DA) improves predictions from these models by aligning them with real-world observations, but marine biogeochemistry DA faces challenges due to model complexity, non-linearity, and sparse, uncertain observations. Existing DA methods applied to marine biogeochemistry struggle to update unobserved variables effectively, while ensemble-based methods are computationally too expensive for high-complexity marine biogeochemistry models. This study demonstrates how machine learning (ML) can improve marine biogeochemistry DA by learning statistical relationships between observed and unobserved variables. We integrate ML-driven balancing schemes into a 1D prototype of a system used to forecast marine biogeochemistry in the North-West European Shelf seas. ML is applied to estimate (i) state-dependent correlations from free-run ensembles and (ii), in an “end-to-end” fashion, analysis increments from an Ensemble Kalman Filter. Our results show that ML improves updates for previously not-updated variables when compared to univariate schemes akin to those used operationally, particularly in lead times smaller than 5 days. Furthermore, ML models exhibit some potential for transferability to new locations, a crucial step toward scaling these methods to 3D operational systems. We conclude that ML offers a clear pathway to overcome current computational bottlenecks in marine biogeochemistry DA and that refining transferability, optimising training data sampling, and evaluating scalability for large-scale marine forecasting, should be future research priorities.

## 1 Introduction

Marine biogeochemistry (BGC) modelling is an essential tool for understanding global marine elemental cycles (e.g., for carbon and nitrogen), as well as for understanding the response of marine ecosystems to a range of human and climate pressures (Heinze and Gehlen, 2013; Ford et al., 2018; Fennel et al., 2022). These pressures include ocean acidification, marine heat waves, and nutrient pollution which can lead to a range of consequences, such as deoxygenation, toxic algal blooms and biodiversity loss (Doney et al., 2009; Smith and Schindler, 2009; Schmidtko et al., 2017; Frölicher and Laufkötter, 2018; Fennel and Testa, 2019; Gobler, 2020). Marine BGC modelling could then support management, policy and planning across a wide range of temporal scales. Marine BGC models are often constrained by the available observations through data assimilation (DA) (Ford et al., 2018; Fennel et al., 2019), providing both multi-decadal reanalyses of past ecosystem trends and variability, as well as short-term operational forecasts (on the scale up to 5-10 days). Such operational forecasts are run by marine forecasting centres in many countries, e.g., by the Copernicus Marine Service in Europe covering the global ocean and all the major European seas (Le Traon et al., 2019).

However, marine BGC DA faces multiple specific challenges (Dowd et al., 2014; Ford et al., 2018; Fennel et al., 2019), compared to assimilation of ocean physics observations in marine models. Marine

48 BGC models are typically more complex than physical models (a pelagic model can have tens of  
49 variables and hundreds of parameters), they are highly non-linear and relatively poorly constrained  
50 (e.g., having highly uncertain parameters) when compared to ocean physics models. Furthermore,  
51 marine BGC observations are even fewer, sparser, and more uncertain than physics observations. This  
52 brings several specific challenges for marine BGC DA, one of those being the need for multivariate  
53 DA, where a large portion of the marine BGC model state variables is updated by observations of only  
54 a small fraction of the model variables. In the context of operational marine BGC forecasting, these  
55 observations are typically satellite ocean colour-derived chlorophyll (Fennel et al., 2019; Groom et al.,  
56 2019). Furthermore the assimilation of BGC-Argo observations (including chlorophyll, nitrate and  
57 oxygen) in open ocean waters has been recently implemented in state-of-the-art operational systems  
58 (Cossarini et al., 2019; Teruzzi et al., 2021). Other products are assimilated in reanalyses or research  
59 and development (R&D) versions of the operational systems, such as optical variables (Shulman et al.,  
60 2013; Ciavatta et al., 2014; Jones et al., 2016; Gregg and Rousseaux, 2017; Skakala et al., 2020) and  
61 size-class chlorophyll (Ciavatta et al., 2018, 2019; Skákala et al., 2018; Pradhan et al., 2020), as well as  
62 types of in situ data, such as chlorophyll, oxygen and nutrients from gliders (Skákala et al., 2021). For  
63 a broader range of marine BGC DA work beyond operational applications, see many other references,  
64 e.g., Simon and Bertino (2012); Shulman et al. (2013); Gehlen et al. (2015); Simon et al. (2015).

65 Different DA systems are used across marine BGC forecasting centres, including variational (Ford  
66 et al., 2012; Song et al., 2016; Skákala et al., 2018; Coppini et al., 2021), Singular Evolutive Extended  
67 Kalman filter (SEEK) (Gutknecht et al., 2019; Ciliberti et al., 2021) and Ensemble Kalman Filter  
68 (EnKF) (Bertino et al., 2021) -based methods. Although ensemble methods (e.g., EnKF) are appeal-  
69 ing for their capability to provide uncertainty quantification and cross-covariances, the more complex  
70 marine BGC models such as the European Regional Seas Ecosystem Model (ERSEM) (Butenschön  
71 et al., 2016) or the Biogeochemical Flux Model (BFM) (Cossarini et al., 2017), currently rely on  
72 variational methods as running a sufficiently large ensemble in the day-to-day operational forecasting  
73 context can be computationally expensive. Moreover, for such complex models, variational methods  
74 update only a very limited number of unobserved variables, typically using very simple balancing prin-  
75 ciples based on the simulated structure and stoichiometry of the phytoplankton community (Teruzzi  
76 et al., 2014; Skákala et al., 2018). We will call such systems with certain approximations “univariate”,  
77 and systems that update (nearly) all model state variables as a direct result of DA “multivariate”.  
78 The multivariate updates can happen in the DA step, through ensemble-informed background<sup>1</sup> error  
79 covariances (as in the EnKF), through balancing schemes, such as the scheme of Hemmings et al.  
80 (2008) based on nitrogen mass conservation applied to Nutrient-Phytoplankton-Zooplankton-Detritus  
81 models (Hemmings et al., 2008; Ford et al., 2012), or through the tangent-linear and adjoint models  
82 Mattern et al. (2017). However, whenever such multivariate schemes were applied to highly com-  
83 plex marine BGC models (in reanalyses, or R&D), the improvement of non-observed variables was  
84 typically marginal, with several variables often systematically degraded by DA (e.g., Ciavatta et al.  
85 (2016, 2018)). This provides a warning on the use of incorrect assumptions in multivariate balancing  
86 schemes or in the EnKFs and the need for better DA and/or ensemble design.

87 The field of machine learning (ML) has developed rapidly during the past few decades, and has  
88 seemingly found function across every level of science and culture, due to the increasing size and  
89 availability of datasets and computational power, together with the continued development of algo-  
90 rithms and theory (Jordan and Mitchell, 2015; Sonnewald et al., 2021). Within Earth sciences, the  
91 flexibility of ML paradigms has allowed its use in a huge variety of applications (Reichstein et al.,  
92 2019), including extensive use in physical ocean modelling (van der Merwe et al., 2007; Nowack et al.,  
93 2018; Kochkov et al., 2021). However, using ML for marine BGC models is comparatively infrequent,  
94 with the most common examples found in parameter estimation (Mattern et al., 2012; Leeds et al.,  
95 2013; Mattern et al., 2014; Schartau et al., 2017). There are only relatively few applications outside  
96 this domain such as using a statistical emulator to quantify uncertainty (Mattern et al., 2013) and  
97 the prediction of hypoxia in shelf sea environments (Skakala et al., 2023).

98 In this work, we investigate whether ML can capture the complex, non-linear relationships among  
99 biogeochemical (BGC) variables, and how these relationships depend on the evolving state of the  
100 system (i.e., are flow-dependent). Here, flow-dependent indicates that the statistical relationships or  
101 errors between variables vary with the current flow or dynamical state, rather than being constant or  
102 solely climatological. The relationships learned by the ML model are then used within a DA scheme,  
103 or as a full substitute for it. Thus, we are not attempting to directly emulate or improve BGC models  
104 via ML, but instead use ML to improve DA, and specifically to cope with the challenging problem  
105 of propagating information from observed to unobserved variables in single-model deterministic runs

---

<sup>1</sup>In data assimilation, the terms “forecast” and “background” are often used interchangeably. Strictly, the background refers to the forecast state used as the prior in the assimilation step.

106 that, by definition, do not have an ensemble to estimate the necessary statistics. The main goal of  
107 this approach is to introduce multivariate DA into the system, whilst benefiting from the relatively  
108 low computational cost of ML. This study falls within a stream of research aimed at building suitable  
109 hybrid ML-DA schemes (see Buizza et al., 2022; Cheng et al., 2023, and references therein), and, to  
110 our knowledge, it is the first such attempt in the context of marine BGC.

111 We first use ML to learn flow-dependent correlations that are needed within a DA update step.  
112 This amounts to merging DA and ML, whereby the latter is used to accomplish a task within the DA  
113 process. We demonstrate that such an ML-based multivariate DA is efficient and accurate. As long  
114 as enough suitable data are available for training, ML is able to learn and map complex non-linear  
115 functions for propagating the information from observed to unobserved portions of the system’s state.

116 In a second configuration, instead of merging DA and ML, DA is used to generate a training  
117 dataset, from which ML learns the full DA step in an “end-to-end” emulation (Barth et al., 2020;  
118 Fablet et al., 2021). In this context, “end-to-end” refers to a ML model trained to map inputs (such as  
119 forecast states and observational information) directly to analysis increments, bypassing traditional,  
120 hand-crafted intermediate steps. As such, the ML model learns to predict the analysis increments for  
121 unobserved variables, given the surface background state and the increment for the observed variable.  
122 Recent work by Bocquet et al. (2024) has demonstrated that the entire EnKF analysis step can be  
123 learned in this data-driven, end-to-end manner.

124 Specifically, we intend to answer the following questions: (a) Can we make improvements to the  
125 existing univariate DA scheme by updating a limited set of additional variables in 1D water column  
126 model, with an ML model to estimate correlations or analysis increments? (b) Can these ML models  
127 be extended to effectively update all unobserved pelagic variables? (c) Is the ML model transferable  
128 to a new location after being trained on some other location?

129 Our work has a potentially important application within the North-West European Shelf (NWES)  
130 operational DA system to which it is tailored. Yet we will discuss its generalisation to other compa-  
131 rable systems, applied to spatial domains with similar type of marine BGC dynamics. Based on the  
132 transferability of the ML model, we speculate whether it is feasible to use the ML model trained in  
133 1D on a 3D domain and propose a methodology for doing so.

134 The paper is structured as follows. We first give, in Sect. 2, details on the 1D physical model, the  
135 BGC model and the configuration used. Also, we establish the setups for the DA workflow, describing  
136 the reference univariate scheme (RUS), and the use of the EnKF. Then, in Sect. 3, we outline the two  
137 ML approaches explored in this work. We also give detail on the ML architecture and climatological  
138 statistics. Next, in Sect. 4, we present and discuss our results for: updating nitrate only; updating  
139 the entire set of pelagic BGC surface variables; and testing the transferability of the ML model to a  
140 new location with different BGC behaviour. In Sect. 5, we draw concluding remarks, summarise the  
141 key findings and discuss future work.

## 142 **2 Model and data assimilation setups for biogeochemistry**

### 143 **2.1 Physical model: GOTM**

144 The Generalised Ocean Turbulence Model (GOTM) (Bolding and Villarreal, 1999) is a 1D water  
145 column model for studying hydrodynamic and biogeochemical processes when coupled to a biogeo-  
146 chemical model, in marine and limnic waters. GOTM provides the necessary physical forcing for  
147 the coupled biogeochemical model, offering a balance between realism and computational cost by  
148 using real atmospheric forcing data, relaxation profiles, and coupling at full BGC complexity, while  
149 sacrificing the explicit representation of 3D processes. This makes the system ideal for this work,  
150 where we are primarily interested in the error relationships between different biogeochemical quan-  
151 tities (e.g., chlorophyll to nitrate), rather than the spatial error characteristics. GOTM can be used  
152 as a stand-alone model for studying dynamics of boundary layers in natural waters, having hydrody-  
153 namic applications in investigations of air-sea fluxes (Vagle et al., 2010), surface mixed-layer dynamics  
154 (Sonntag and Hense, 2011), dynamics of bottom boundary layers with or without sediment transport  
155 (Umlauf and Burchard, 2011; Falchetti et al., 2010), and estuarine and coastal dynamics (Burchard,  
156 2009).

### 157 **2.2 Biogeochemical model: ERSEM**

158 ERSEM (Baretta et al., 1995; Butenschön et al., 2016) is a marine biogeochemistry model that sim-  
159 ulates lower trophic levels of the ocean ecosystem, including plankton and benthic fauna (Blackford,

160 1997), see Table 1. The model divides phytoplankton into four functional types based on size: pico-  
 161 phytoplankton, nanophytoplankton, microphytoplankton and diatoms (Baretta et al., 1995). ERSEM  
 162 uses variable stoichiometry for the simulated plankton groups (Baretta-Bekker et al., 1997; Geider  
 163 et al., 1997) and represents the biomass of each functional type in terms of chlorophyll, carbon, ni-  
 164 trogen, and phosphorus, with diatoms also being represented by silicon. ERSEM predators consist of  
 165 three types of zooplankton (mesozooplankton, microzooplankton, and heterotrophic nanoflagellates),  
 166 with organic material being decomposed by a single type of heterotrophic bacteria (Butenschön et al.,  
 167 2016). The model represents three different sizes of detritus (small, medium and large) and three  
 168 types of dissolved organic matter (DOM: refractory; semi-labile; labile). The inorganic component of  
 169 ERSEM includes nutrients such as nitrate, phosphate, silicate, ammonium, and carbon, as well as dis-  
 170 solved oxygen. The carbonate system is also included in the model (Artioli et al., 2012). ERSEM has  
 171 been used for many applications including NWES and Mediterranean Sea biogeochemistry reanalyses  
 172 (Ciavatta et al., 2016, 2018, 2019), NWES operational forecasts (Skákala et al., 2018; McEwan et al.,  
 173 2021), and NWES climate projections (Wakelin et al., 2015, 2020; Galli et al., 2024).

Functional Group	Class/Type	Chemical Components
Phytoplankton	Diatoms	$\chi_{\text{dia}}$ , C, N, P, Si
Functional Types (PFT)	Microphytoplankton	$\chi_{\text{micro}}$ , C, N, P
	Nanophytoplankton	$\chi_{\text{nano}}$ , C, N, P
	Picophytoplankton	$\chi_{\text{pico}}$ , C, N, P
Zooplankton	Mesozooplankton	C
	Microzooplankton	C, N, P
	Heterotrophic Flagellates	C, N, P
Bacteria	—	C, N, P
Detritus	Small	C, N, P
	Medium	C, N, P, Si
	Large	C, N, P, Si
Dissolved Organic Matter (DOM)	Labile	C, N, P
	Semi-labile	C
	Refractory	C
Nutrient	Nitrate ( $NO_3^-$ )	N
	Phosphate ( $PO_4^{3-}$ )	P
	Ammonium ( $NH_4^+$ )	N
	Silicate ( $SiO_4^{4-}$ )	Si
Other	Temperature	—
	Oxygen $O_2$	—

Table 1: Reference table for ERSEM pelagic variables used in this study. Chemical components are represented by the following symbols:  $\chi$  is chlorophyll; C is carbon; N is nitrogen; P is phosphorus and Si is silicon. Note that we also use total chlorophyll (denoted as chl in this work), which is a diagnostic variable calculated as the sum of chlorophyll concentrations from all PFT classes.

174 The coupler known as the “Framework for Aquatic Biogeochemical Models” (FABM) (Bruggeman  
 175 and Bolding, 2014) allows for the smooth combination of hydrodynamic and biogeochemical models,  
 176 and is used to couple GOTM with ERSEM in this work. The coupling of GOTM to marine BGC  
 177 models using FABM has allowed for a wide range of applications that include modelling of phytoplank-  
 178 ton growth (Kerimoglu et al., 2021), examining the implications of sea-ice BGC for oceanic emissions  
 179 (Hayashida et al., 2017), assessing the highly intermittent spatial variability of phytoplankton on sub-  
 180 grid scales (Mandal et al., 2016), and enhancing stoichiometry in existing BGC models (Anugerahanti  
 181 et al., 2021).

### 182 2.3 Model configuration and synthetic data setup

183 We configure the GOTM-FABM-ERSEM setup for two different locations in the English Channel (see  
 184 Fig. 1) and use synthetic observations of each. The first location, known as L4 (50.25°N, 4.217°W),  
 185 is a highly biologically productive site with seasonally stratified dynamics (Pingree and Griffiths,  
 186 1978), influenced significantly by the outflow of the nearby Tamar and Plym rivers. Nitrate acts as  
 187 the primary limiting nutrient for phytoplankton growth. It is monitored by the Western Channel  
 188 Observatory (WCO) (<https://www.westernchannelobservatory.org.uk/>) and SmartSound Plymouth  
 189 (<https://www.smartsoundplymouth.co.uk/>).

190 Besides the L4 site, we configure a setup for an additional location, that we shall refer to as the  
 191 Central Western English Channel (CWEC), at 49.40°N, 4.217°W. This point is less biologically pro-  
 192 ductive and it is much less influenced by riverine outflow than L4. These differences are evident when  
 193 looking at the distributions of biogeochemical signals in the models applied at these two locations (see  
 194 Fig. A.1). The differences make CWEC a reasonable alternative test site for assessing the application  
 195 of the ML model, and its suitability to generalise the results of this study under different marine BGC  
 196 conditions.

197 The physical and biogeochemical models for each location are forced with data appropriate for the  
 198 study area, using the following datasets: the General Bathymetric Chart of the Oceans 2023 (1/240°  
 199 resolution) for water depth; the ECMWF ERA5 dataset (0.25°/hourly resolution) for meteorology;  
 200 the TPXO9-atlas (1/30° resolution) for tides; and the World Ocean Atlas 2018 (0.25° resolution) for  
 201 temperature, salinity and nutrient fields (nitrate, phosphate and silicate) for biogeochemical relaxation  
 202 profiles. A nutrient relaxation timescale of 3 months towards the World Ocean Atlas data is required  
 203 to prevent significant trends forming that cause the 1D model to gradually accumulate nutrients.  
 204 This relaxation is significantly longer than the assimilation cycle of 7 days, and so has little impact  
 205 on forecast errors at the surface. However, the relaxation profiles could contribute to controlling  
 206 behaviour below the mixed layer in our setup (which is not updated during assimilation). This could  
 207 help to mitigate some long-term biases in these areas. This is a potential limitation for operational  
 208 scale systems which do not have or use these relaxation profiles.

209 Ensemble runs, whether as free runs or for the EnKF (see Sect. 2.4.2) are configured and run using  
 210 the Ensemble and Assimilation Tool (EAT) in Python (Bruggeman et al., 2024). Each ensemble is  
 211 given a spin-up period of 10 years to settle the biogeochemistry appropriately and provide well-spread  
 212 initial conditions. Each ensemble member is perturbed by temporally correlated random noise to  
 213 scale the ECMWF ERA wind forcing at the location. This is used to scale the wind forcing and has a  
 214 correlation timescale of 7 days, a mean of 1 and a standard deviation of 0.5. The resulting variation in  
 215 wind strength across the ensemble members increases their spread over time, and prevents ensemble  
 216 collapse (at least within the mixed layer) induced by the previously mentioned nutrient relaxation,  
 217 or lack of representation of error growth processes like horizontal advection which are absent in a 1D  
 218 set-up.

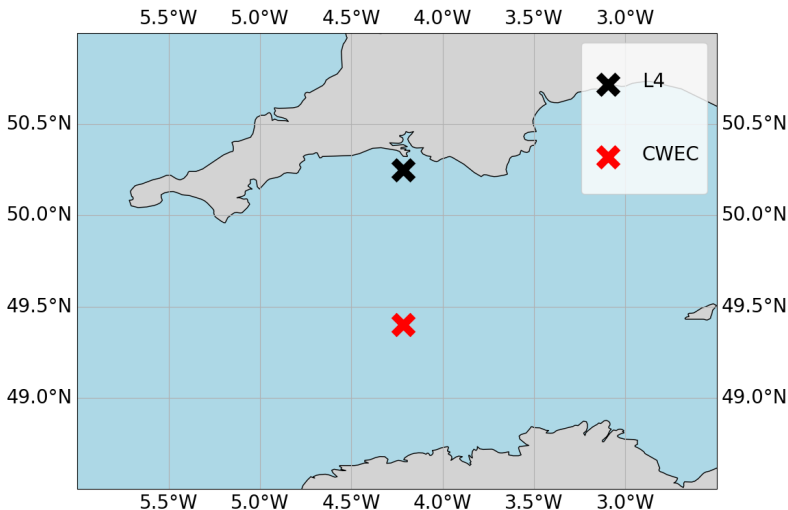


Figure 1: Map of the Western English Channel, marking the L4 model-training location with a black cross and the CWEC (Central Western English Channel) with a red cross, where we evaluated the model portability.

219 For the purposes of training ML models, and generating climatological statistics, time periods for  
 220 the L4 location are partitioned as follows: training data (2000-2014), validation (2015-2017), offline  
 221 test (2018-2020), online test (2022-2023). Offline refers here to a setup in which the DA analysis is  
 222 not then cycled as the initial condition for the next forecast, and so it does not impact successive DA  
 223 cycles. Online refers to a setup in which assimilation updates are cycled, and so can have dynamical  
 224 impact on later DA cycles as the model integrates in time. Climatological correlations and variances

225 are calculated using a free-run ensemble over the training period. Because the forecasts extends seven  
 226 days into the future, the number of available samples for any specific calendar day was limited, even  
 227 though forecasts were issued throughout the full training period. To address this, the climatological  
 228 statistics were computed as daily values, defined by averaging the statistics for a given day across all  
 229 years in the training set. To increase the sample size and smooth out variability, a  $\pm 30$ -day window  
 230 around each calendar day was applied. The CWEC location uses a run spanning 2000 - 2010 to  
 231 generate climatological statistics, and the online test is performed for (2022-2023). No validation  
 232 or offline test period is required for CWEC in this work, as we are primarily interested in a “naive  
 233 transferability” of the model trained and evaluated at L4.

## 234 2.4 Data assimilation setups

235 We examine five data assimilation (DA) setups in this work: a simple univariate scheme representative  
 236 of current operational marine BGC systems, a scheme that uses climatological statistics to update  
 237 unobserved variables, an EnKF for comparison, and two novel hybrid DA schemes which are hybridised  
 238 with ML techniques. Before describing each scheme, we give the basic equations of conventional DA,  
 239 introduce the state vectors that the DA uses, the observation type that we assimilate, and other  
 240 adjustments that are done post assimilation.

241 The update equation that is central to DA is sometimes called the best linear unbiased estimator  
 242 (BLUE, Asch et al., 2016; Carrassi et al., 2018) and is given by

$$\mathbf{x}^a = \mathbf{x}^b + \underbrace{\mathbf{K}(\mathbf{y} - \mathcal{H}(\mathbf{x}^b))}_{\text{analysis increment}}, \quad (1)$$

243 where  $\mathbf{K}$  is the Kalman gain matrix

$$\mathbf{K} = \mathbf{P}^b \mathbf{H}^\top (\mathbf{H} \mathbf{P}^b \mathbf{H}^\top + \mathbf{R})^{-1}, \quad (2)$$

244  $\mathbf{x}^a$  is the analysis (updated) state,  $\mathbf{x}^b$  is the background state (which in this work is the state of  
 245 the system after a seven-day forecast period),  $\mathbf{y}$  are the observations,  $\mathcal{H}$  is the observation operator  
 246 (with Jacobian  $\mathbf{H}$ ),  $\mathbf{P}^b$  is the background error covariance matrix, and  $\mathbf{R}$  is the observation error  
 247 covariance matrix. The matrix  $\mathbf{P}^b$  is of special interest to this work. Ideally this matrix should be  
 248 appropriately flow-dependent, but in practice it is often not, such as in many operational schemes  
 249 where a fixed climatological estimate is used. For instance, in marine BGC, we would expect the  
 250 onset of a phytoplankton bloom to trigger changes in nutrient concentrations, which in turn affect  
 251 the correlations between the two. This response cannot be captured by climatological statistics alone.  
 252 The purpose of this work is to introduce such flow-dependency to  $\mathbf{P}^b$ , or to the analysis increments  
 253  $\Delta \mathbf{x}$ , with ML techniques. While each of the DA schemes described later will use different methods to  
 254 update unobserved variables, all schemes will update total chlorophyll, the observed variable, and its  
 255 associated PFTs in the same way (using the RUS scheme described in Sect. 2.4.1).

256 In this work, the state vectors of the DA system,  $\mathbf{x}^b$  and  $\mathbf{x}^a$ , consists of the surface values of total  
 257 chlorophyll and a set of chosen unobserved pelagic ERSEM variables (with chosen variable setups  
 258 detailed in Sect. 3). While a state vector only considers surface values (0D), the entire DA process  
 259 itself is 1D, using an idealised structure function to spread increments uniformly throughout the mixed  
 260 layer. The mixed layer consists of a number of vertical GOTM grid cells, which can vary throughout  
 261 the year based on the physical drivers determining mixed layer depth, such as temperature and salin-  
 262 ity. This approach assumes that surface conditions are representative of the mixed layer as a whole  
 263 (which is broadly true in this model configuration, where vertical mixing is strong enough to ensure  
 264 that surface and mixed layer conditions remain well-coupled). As a result, the increments derived at  
 265 the surface also provide a reasonable correction at depths/other grid cells within the mixed layer. We  
 266 do not update the variables below the mixed layer, as they are decoupled or weakly coupled with the  
 267 surface. Extending increments deeper would risk introducing spurious vertical structures, distorting  
 268 stratification, or interfering with biogeochemical processes that are driven by different controls (e.g.,  
 269 remineralisation). By restricting updates to the mixed layer, the assimilation scheme ensures consis-  
 270 tency with the available observations and avoids imposing unsupported corrections in the sub-mixed  
 271 layer. Strategies for updating below the mixed layer may therefore require additional observation  
 272 types (e.g., from floats or sea-gliders) or bias-correction approaches, rather than relying solely on  
 273 surface observations combined with DA.

274 We assimilate only observations of total chlorophyll,  $y$ , at the surface. Total chlorophyll in the  
 275 model,  $x_{\text{chl}}$ , is a diagnostic variable obtained by summing the chlorophyll content of all phytoplank-  
 276 ton functional types (PFTs, see Table 1), which are themselves prognostic variables. In principle,

277 one could keep only the PFT chlorophyll concentrations in the state and represent total chlorophyll  
 278 via the observation operator. However, this would require explicitly summing across PFTs at every  
 279 assimilation step, making the DA equations more cumbersome. Instead, we treat surface total chloro-  
 280 phyll directly as part of the state vector. This simplifies the presentation of the analysis equations,  
 281 since the observation operator then reduces to a simple selection operator:

$$\mathbf{H} = [1, 0_1, \dots, 0_N], \quad (3)$$

282 with the state ordered as  $(x_{chl}, x_1, \dots, x_N)$ . The system dimension is therefore  $N + 1$ , with the first  
 283 element corresponding to surface total chlorophyll, and the remaining elements corresponding to the  
 284 unobserved variables to be updated.  $\mathbf{H}$  is a row vector, as there is only one observation per DA  
 285 update.

286 The individual PFTs themselves are excluded from the state, and instead updated after the main  
 287 analysis step in Eq. (1). The surface total chlorophyll increment is redistributed across PFTs in  
 288 proportion to their background contribution to total chlorophyll  $x_{chl}$ :

$$x_{\chi_i}^a = x_{\chi_i}^b + \frac{x_{\chi_i}^b}{x_{chl}^b} \cdot (x_{chl}^a - x_{chl}^b), \quad (4)$$

289 where  $\chi_i$  stands for the chlorophyll component of each PFT,  $i$ , as in Table 1.

290 The associated chemical components of each PFT (C, N, P, and for diatoms also Si) are then  
 291 updated in proportion to their background stoichiometric ratios with PFT chlorophyll:

$$x_{\zeta_{i,j}}^a = x_{\zeta_{i,j}}^b + \frac{x_{\zeta_{i,j}}^b}{x_{\chi_i}^b} \cdot (x_{\chi_i}^a - x_{\chi_i}^b), \quad (5)$$

292 where  $\zeta_{i,j}$  represents the non-chlorophyll components,  $j$ , of each PFT,  $i$ . This constrained redistribu-  
 293 tion scheme (Teruzzi et al., 2014; Skákala et al., 2018) ensures that phytoplankton updates preserve  
 294 forecast stoichiometric ratios and remain physiologically consistent, rather than allowing a statistical  
 295 method (such as the EnKF) to update PFTs freely. While we would expect an unconstrained EnKF  
 296 to eventually converge towards similar balances, this explicit approach guarantees that assimilation  
 297 respects acclimation dynamics and maintains the community structure of the model. Importantly, the  
 298 same ratio-based balancing scheme can also be applied outside an ensemble framework in single-model  
 299 runs, where it provides a consistent way of updating PFTs from a total chlorophyll correction.

300 The above observation operator, and updates to the PFTs are used in all DA schemes in this  
 301 paper. The specific DA schemes (conventional and ML-based) are now described. A summary of the  
 302 methods is given in Table 2.

### 303 2.4.1 Reference univariate DA scheme (RUS)

304 We call our baseline DA method the reference “univariate” DA scheme (RUS, Table 2, row 4). Its  
 305 purpose is to mimic existing DA systems used by several operational centres (Teruzzi et al., 2014;  
 306 Skákala et al., 2018), although our scheme is not variational. The background error covariances are  
 307 based on climatological information and so do not adapt to the state.

308 The RUS is based on an evaluation of Eq. (1), but only to directly update the total chlorophyll  
 309 variable. The simple structure of the observation operator in Eq. (3) means we can rewrite the update  
 310 Eq. (1) to show how the total chlorophyll (index *chl*) is updated from the total chlorophyll observation:

$$x_{chl}^a = x_{chl}^b + \frac{P_{chl,chl}^b}{P_{chl,chl}^b + R} \cdot (y - x_{chl}^b), \quad (6)$$

311 where  $P_{chl,chl}^b$  is the background error variance of total chlorophyll and  $R$  is the observation error  
 312 variance. Climatological variances, which are used to estimate  $P_{chl,chl}^b$ , are calculated from the same  
 313 period of the EnKF run (Sect. 2.4.2) used to train the ML models. Details on the training runs can  
 314 be found in Sect. 3.1 and 3.2. Updates to the surface PFT chlorophyll, to the associated chemical  
 315 components, and throughout the mixed layer are made separately as described previously in Sect. 2.4.

316 Note that we call this scheme “univariate” as only a single variable (total chlorophyll) is updated  
 317 according to the background and observational errors as described in Eq. (6). All further DA schemes  
 318 described in this work (apart from the EnKF below, which uses ensemble-derived covariances) start  
 319 with an update of the total chlorophyll using Eq. (6), and will attempt to update additional pelagic  
 320 variables using the new ML-based approaches.

Run / scheme	Description / purpose	<i>chl</i> variance	<i>i</i> variance	<i>chl-i</i> correlation source	$\Delta x_{chl}$	$\Delta x_i$
<b>Preparation / training runs</b>						
1. Truth run	To synthesise observations and for analysis evaluation	n/a	n/a	n/a	n/a	n/a
2. Ensemble of free-runs	To determine climatological correlations and training for ML-OI	n/a	n/a	n/a	n/a	n/a
3. EnKF	Update all chosen surface variables / gold standard run / training for ML-EtE	ensemble-based	ensemble-based	ensemble-based	Eq. (1)	Eq. (1)
<b>Conventional assimilation runs</b>						
4. RUS	Reference univariate scheme (TC + stoichiometrical PFT update) / baseline for extensions	climatology	n/a	n/a	Eq. (6)	zero
<b>RUS extension assimilation runs (update to variable <i>i</i> with ML methods)</b>						
5. ML-OI	ML correlation hybrid	climatology	climatology	ML of free run	As RUS	Eq. (7)
6. ML-EtE	ML end-to-end EnKF emulation	climatology	n/a	n/a	As RUS	ML
<b>RUS extension assimilation runs (update to variable <i>i</i> with non-ML methods)</b>						
7. CliC	Climatological correlations	climatology	climatology	climatology	As RUS	Eq. (7)

Table 2: An overview of the different run-types and schemes used in this work. Index *chl* refers to total chlorophyll, while index *i* refers to an unobserved variable (e.g., nitrate). The truth run refers to a single-model run with no updates. We sample synthetic observations from this run and feed these into each DA scheme. The ensemble of free-runs means the model is left to run without assimilation. The EnKF uses an ensemble to model background error covariance in the DA update of all state variables (Sect. 2.4.2). The RUS is the ‘univariate’ scheme (Sect. 2.4.1), which is used as a benchmark for the performance of other schemes. It updates only the total chlorophyll state variable. The ML-OI estimates background correlations of variables beyond the total chlorophyll with an ANN (Sect. 3.1). The ML-EtE estimates the analysis increments of variables beyond the total chlorophyll produced by an EnKF using an ANN (Sect. 3.2). The CliC is similar to ML-OI but uses purely climatological background statistical estimates of the correlations to update the state of unobserved variables (Sect. 3.3).

#### 2.4.2 The EnKF-based scheme

The stochastic EnKF scheme, see e.g., Evensen (2003), approximates the update Eqs. (1) and (2) with an ensemble to estimate the flow-dependent background error covariance matrix  $\mathbf{P}^b$ , Table 2, row 3. For each ensemble member there is a different update and a different perturbed observation (the perturbations are sampled from the normal distribution  $\mathcal{N}(0, R)$ ). The EnKF updates all elements of the surface state described previously using the ensemble version of Eq. (1), but still performs the stoichiometric balancing scheme and duplication of the analysis increments from the surface throughout the mixed layer, as described in Sect. 2.4. This is done so that there is a one-to-one correspondence between the strategy used to generate the training data, and the strategy applied in the single-model schemes.

### 3 Hybrid machine learning data assimilation for marine biogeochemistry

In this section, we describe how we hybridise the DA, described above, with ML to provide flow-dependent estimates of the statistics/increments that are better than the climatological values. In particular, we take two approaches that differently replace parts of, or fully, the update equation. We now show the mathematical framework that the ML schemes will emulate, which is derived from Eqs. (1)-(3).

The ML-based DA schemes are summarised in Table 2, rows 5 and 6. They both build upon RUS, extending it to become multivariate. The total chlorophyll analysis is computed using the RUS update Eq. (6), while the remaining variables ( $1 \leq i \leq N$ ) have updates according to

$$x_i^a = x_i^b + \underbrace{\frac{P_{i,chl}^b}{P_{chl,chl}^b + R}}_{\text{analysis increment}} \cdot (y - x_{chl}^b), \quad (7)$$

where  $P_{i,chl}^b$  is the background error covariance between variable  $i$  and total chlorophyll defined as

$$P_{i,chl}^b = \rho_{i,chl} \cdot \sigma_i \cdot \sigma_{chl}, \quad (8)$$

where  $\rho_{i,chl}$  is their background error correlation, and  $\sigma_i$  and  $\sigma_{chl}$  are their respective background error standard deviations. In Eq. (7) the analysis increment of the update is labelled.

An important aspect of any DA scheme is its ability to adapt with the flow. A conventional way to introduce flow-dependency is via Monte Carlo-like methods such the EnKF, which comes with substantial computational cost. The two proposed ML-DA schemes below are designed with the above in mind and provide flow-dependency cost-effectively without the need for an ensemble (apart from at the training stage, as shall be clarified). The two ML-DA schemes are described in Sections 3.1 and 3.2.

Regardless of the specific ML-DA scheme, each ML model is a fully connected artificial neural network (ANN) optimized using AutoKeras (Jin et al., 2019) over 100 trial configurations. AutoKeras uses Bayesian optimisation in a network search algorithm to determine optimal hyperparameters such as layer depth, layer width, dropout rate, learning rate, and optimiser selection. The input features of each model are standardised to have unit variance and a mean of zero, using data from the L4 training period during training.

Each ML approach is tested in the following scenarios: (1) a set-up where the only unobserved variable that is updated is nitrate, (2a) a set-up where we update the full set of pelagic variables, and (2b) a set-up where we update a partial set of the pelagic variables, eliminating poorly estimated variables based on the results of (2a). The progression from (1) to (2a) allows us to move from a controlled test of a single key limiting variable to a comprehensive update of the full system, while (2b) represents a refinement step that is only possible after evaluating the performance of (2a). In this way, the experimental design not only tests the limits of updating all variables, but also demonstrates how excluding problematic variables can improve robustness without discarding the broader benefits of multivariate updates. In each setup, the number of outputs to be estimated by the ML model corresponds to the number of unobserved variables. In the case of ML-OI (see Sect. 3.1), the standard deviations must also be estimated from climatology.

#### 3.1 Hybrid machine-learning optimal interpolation (ML-OI)

This approach first updates the observed total chlorophyll and associated PFTs in an identical manner to the RUS described in Sect. 2.4.1. Then, an ANN estimates the state-dependent correlations  $\rho_{i,chl}$  between observed and unobserved quantities in Eq. (8) as a function of the background state (even though the EnKF update of state variables from other state variables is linear, the relationship between state variables and correlations is likely non-linear). Together with climatologically estimated values of  $\sigma_i$  and  $\sigma_{chl}$  (estimated using a free-run ensemble over the training period as described in Sec. 2.3), the correlations are substituted into Eqs. (8) and then (7) to provide updates to the unobserved variables in the system. We call this approach ML-OI (“optimal interpolation”, Table 2, row 5). For each variable input into the ML-OI model to estimate the correlation, the background state is additionally divided by its climatological maximum at the corresponding location (before the regular standardisation procedure is applied to the data). This normalisation accounts for differences in the

379 amplitude of seasonal variability while making an assumption that the underlying correlative relation-  
380 ships between variables remain consistent across locations. Since correlations are dimensionless, this  
381 scaling does not affect their interpretation, and the subsequent analysis increments (which are con-  
382 structed by combining the estimated correlations with the location-specific climatological variances)  
383 remain physically consistent. Apart from correlation, the other aspect of covariance to determine are  
384 the variances. Since each variance is a single-variable statistic that can be estimated more reliably  
385 than correlations from long climatological records, we assume they are sufficiently robust to provide a  
386 stable basis for use in the Kalman gain. In contrast, correlations describe joint variability and therefore  
387 require the more sophisticated, data-driven approach described above, and cannot be approximated  
388 in the same way.

389 As with every approach used in this work, the resulting surface increments of the unobserved  
390 variables are then propagated to the other levels in the mixed layer, as described previously.

391 In order to generate training data for this approach, we run a 100-member ensemble of free-  
392 runs, configured according to Sect. 2.3 (Table 2, row 2). We generate training samples at seven day  
393 intervals across these free-runs, covering the period from 2000-2014. The features are the surface  
394 states of individual ensemble members at a given time, across all pelagic model variables. For the  
395 first application of ML-OI in Sect. 4.2, the targets are time dependent/ensemble-derived correlations  
396 between total chlorophyll and nitrate. In the later application in Sect. 4.3 onwards this is extended  
397 from just nitrate to a wider set of variables.

398 While the field of hybrid ML-DA is growing rapidly, there exists relatively few works in which ML-  
399 estimated background error covariances are so closely coupled with existing DA systems. However,  
400 a few particularly relevant examples stand out such as Ouala et al. (2018), in which a Kalman-like  
401 analysis update is applied to satellite-derived sea surface temperature fields using ANN-estimated  
402 background error covariances. Additional examples of this can be seen in Sacco et al. (2022), which  
403 aim to learn different sources of uncertainty using ANNs on both toy models and sea level pressure  
404 forecasts. Further work (Sacco et al., 2024) uses an EnKF to generate flow dependent background  
405 error covariances, and then learns them using a convolutional neural network.

### 406 **3.2 End-to-end machine learning of EnKF updates (ML-EtE)**

407 This approach again first updates the observed total chlorophyll and associated PFTs in an identical  
408 manner to the RUS described in Sect. 2.4.1. Nevertheless, as opposed to ML-OI, ML is used here to  
409 estimate the analysis increments for unobserved variables, given the analysis increment of the observed  
410 variable (total chlorophyll) and the complete background state. This obviously requires running a DA  
411 system to learn from. This is achieved here using the updates produced by an EnKF training run (see  
412 below). We call this approach ML-EtE (“end-to-end”, Table 2, row 6), as it emulates an existing DA  
413 system.

414 In ML-EtE, the analysis increment is predicted directly. By contrast, ML-OI predicts correlations,  
415 which are then combined with climatological standard deviations to form a Kalman gain, and only  
416 then applied to the innovation to yield the increment. The ML-OI scheme introduces potential sources  
417 of error – both from the uncertainty of data-driven correlation estimates and from the reliance on  
418 climatological statistics. Directly estimating the analysis increment (a vector) is also naturally more  
419 scalable for high-dimensional applications (e.g., operational 3D systems), where manipulating the full  
420 error covariance matrix (or even reduced or reformulated versions) becomes computationally costly.

421 To generate the training data for this approach, we first generate a nature run for the training  
422 period (Table 2, row 1), to generate synthetic surface observations of total chlorophyll concentration  
423 at weekly intervals. The observation uncertainty is equal to 10% of the observed value. These are then  
424 assimilated into the EnKF run over the same period. The features of each training sample consist of an  
425 individual ensemble member’s background state and its corresponding total chlorophyll increment from  
426 the EnKF run. The targets are the corresponding analysis increments for the unobserved variables at  
427 the surface. As described previously, the resulting surface increments of the unobserved variables are  
428 then propagated to the other levels within the mixed layer.

429 This approach follows other non-marine BGC work in a similar direction, such as Bonavita and  
430 Laloyaux (2020), who used ANNs to emulate the main features of an operational weak-constraint  
431 4D-Var scheme, while Bocquet et al. (2024) pursued a similar end-to-end replacement of the analysis  
432 step. Likewise, several studies have demonstrated the emulation of analysis increments to estimate and  
433 correct model error (Brajard et al., 2020; Gregory et al., 2024). A common feature of these works is  
434 their reliance on an existing DA system or, more generally, on a robust reanalysis. The need for such a  
435 reanalysis represents a key limitation of this approach – one we revisit in the conclusion. Here, however,  
436 our primary goal is to examine the feasibility of ML-EtE and its ability to learn the EnKF updates

effectively. Using this approach, the increments still ultimately come from the “analysis–background” of an EnKF, which provides a linear update to the system (even if the relationship between the state at the analysis increments is non-linear). Going beyond this limitation would require either training on increments relative to the true state (e.g., truth–background) or employing a non-linear DA system, such as a particle filter, to capture more complex, non-linear corrections.

### 3.3 Purely climatological updates

A further non-ML-based scheme is used to update the nitrate to mirror ML-OI, but using only climatological correlations derived from the EnKF run (CliC, Table 2, row 7) over the training period. This serves as another comparison point, a benchmark, to check whether the additional complexity of an ML model is needed.

### 3.4 Skill metric and machine learning model evaluation

#### 3.4.1 Skill metric

For a system that runs for  $\tau$  cycles (where a cycle represents a complete 7-day forecast and analysis), we represent the trajectory for a member  $i$  of ensemble  $X$  at cycle  $t$  as  $X_t^i$ . The truth is denoted as  $T_t$ . The expected RMSE (root mean square error) over  $M$  ensemble members (or a set of  $M$  single-model runs), is calculated as:

$$\text{RMSE} = \frac{1}{M} \sum_{i=1}^M \sqrt{\frac{1}{\tau} \sum_{t=1}^{\tau} (X_t^i - T_t)^2}. \quad (9)$$

This is a sensible metric to use when calculating the expected error across a set of independent single-model runs, such as in the RUS, ML-OI, ML-EtE, and CliC approaches. It is also convenient for calculating the expected error of the ensemble members in the EnKF runs.

#### 3.4.2 SHAP analysis

Shapley values are a well known and widely used metric for understanding the importance and contributions of individual input features in ML models (Lundberg and Lee, 2017). A Shapley value represents the average marginal contribution of a feature across all possible subsets of features, ensuring a fair allocation of importance. In this work, we use Kernel SHAP (SHapley Additive exPlanations) as a model-agnostic approach to estimate mean Shapley values across a dataset. Kernel SHAP approximates Shapley values by training a weighted linear model on perturbations of the input data. By calculating the mean absolute Shapley values, we measure the magnitude of influence for individual features to the model’s estimations.

By understanding the importance of each input feature, we gain insight into the correlative links of dynamical behaviour in the system. This can help to identify how-and-when a model will translate well to new conditions. For example, if the primary predictive feature of an ML model is similar in two separate locations, one trained and one unseen, then we may expect the ML model to perform reasonably well in the new scenario, even if the other non-predictive features exhibit an entirely different distribution. We emphasise that we cannot infer causality from this analysis alone but understanding the data-driven feature importance and feature contribution for an ML model, combined with expert understanding of the system dynamics, can help to unveil connections and insights into the complex processes of the marine BGC model.

It also worth noting that these metrics can also be used for feature selection with the idea that if a feature contributes little-to-nothing to the predictions, it can probably be eliminated from the feature set. This then requires the expensive processes of iteratively re-training and re-testing the neural networks and so is not an avenue that we explore in this work. SHAP is somewhat limited in the presence of highly correlated features because Shapley values assume independent feature contributions. This can lead to arbitrary or shared attributions when features provide redundant information, making it difficult to disentangle their true individual impacts. However, the correlation structures of the marine BGC have been studied previously (Higgs et al., 2024), and so can be more effectively accounted for during analysis.

## 4 Results and discussion

### 4.1 System dynamics

As discussed in Sect. 2.3, the L4 location is a highly biologically productive site with seasonally stratified dynamics. Nitrogen is a key component of organic matter and is generally the limiting nutrient to primary production by phytoplankton in coastal marine ecosystems (Council et al., 2000), which includes the L4 location (Smyth et al., 2010). This leads to a strong, potentially exploitable dynamical link between phytoplankton and nitrate that varies with a clear seasonal cycle. Figure 2 demonstrates this seasonality, and how it can be broken down into three distinct regimes across any given year:

- The *light-limited regime* approximately spans the period from October to the start of the next spring bloom. Here, there is little-to-no phytoplankton growth due to the reduced light-levels at this time of year, meaning nutrients can be mixed throughout the water column without being used by the phytoplankton. During this period, phytoplankton concentrations are very low and mostly decoupled from nutrient dynamics <sup>2</sup>.
- The *bloom regime* can occur throughout spring (from March until May), and is the period when phytoplankton reaches its yearly maximum. During this time, light levels no longer limit phytoplankton growth and there is a high availability of nutrients that have accumulated in the water column during the “light-limited” period. This results in a rapid increase of phytoplankton concentration, and an exhaustion of nutrients.
- The *nutrient-limited regime* refers to the period roughly spanning from early summer until late September where nutrients, and more specifically nitrate, have been exhausted by the phytoplankton during bloom and so concentrations are generally very low. During this time, phytoplankton relies on processes such as storms to mix nutrients into the upper water column. Consequently, phytoplankton growth is sporadic and less intense than during the spring bloom.

---

<sup>2</sup>The model is constrained to physically non-negative values, so any negative values that arise during the data assimilation (DA) step, though extremely rare, are clipped to zero. This occurs infrequently and only in very localized instances. As such, this treatment has a negligible impact on the overall state and statistical distributions. Additionally, due to strong surface forcing, the model tends to quickly redistribute any localized anomalies, minimizing the persistence or propagation of these clipped values. Therefore, we are confident that this approach does not significantly influence the model performance or results.

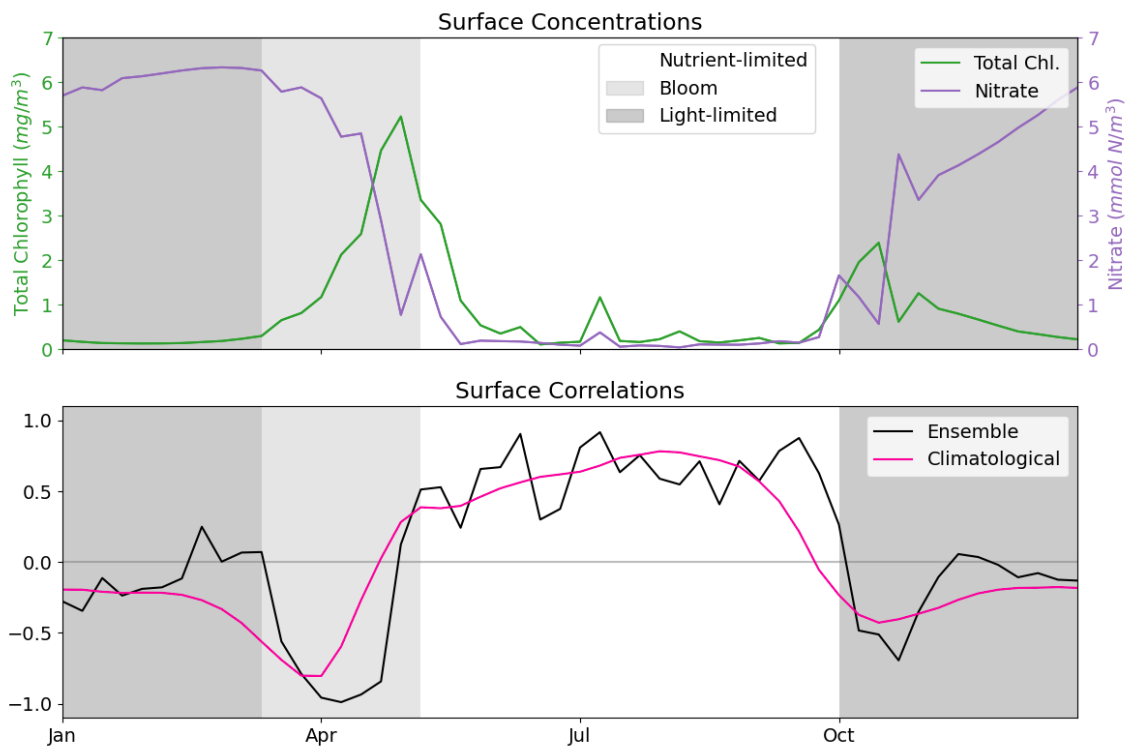


Figure 2: The top panel shows a time series for the surface concentrations of total chlorophyll (green) and nitrate (purple) during 2023 at the L4 location. The bottom panel presents correlations between total chlorophyll and nitrate derived from a 96-member ensemble (black) and from a daily varying climatology (red; calculated over the 2000–2014 training period). Shading indicates the dominant seasonal system regimes: “light-limited” (dark grey), “bloom” (light grey) and “nutrient-limited” (white).

## 4.2 Estimation and update to a single pelagic variable

In this section, we explore the performance of ML-OI and ML-EtE in updating only nitrate as an unobserved variable. Recall however that the observed total chlorophyll and associated PFTs are updated according to the RUS scheme in Sect. 2.4.1. We choose nitrate for these initial experiments because it is a limiting nutrient at the L4 location (see Fig. A.2 in the Appendix and Smyth et al., 2010), and therefore has a clear, explainable relationship with total chlorophyll as discussed in Sect. 4.1 (see also Fig. 2). Since nitrate is the key driver limiting primary production among nutrients, addressing it through DA could have a significant knock-on effect on the whole model state (through dynamical evolution from the corrected state). Moreover, this also provides the simplest proof-of-concept system to analyse initially, before we later extend the updates to more than 30 additional pelagic variables (see Table 1) in a higher complexity scenario. However, as will become clear in Sect. 4.3, this strategy for assigning importance to variables in the DA scheme does not necessarily reflect their true dynamical role in the model, highlighting the need to evaluate how different assimilation strategies and variable-update choices affect performance.

Figure 3 shows the correlation between total chlorophyll and nitrate as a function of time in the period 2018–2020 for the “offline” ML-OI experiment. The performance of ML-OI is compared to the “true correlation” computed over an ensemble of 100 members and to the correlation estimated using daily climatology.

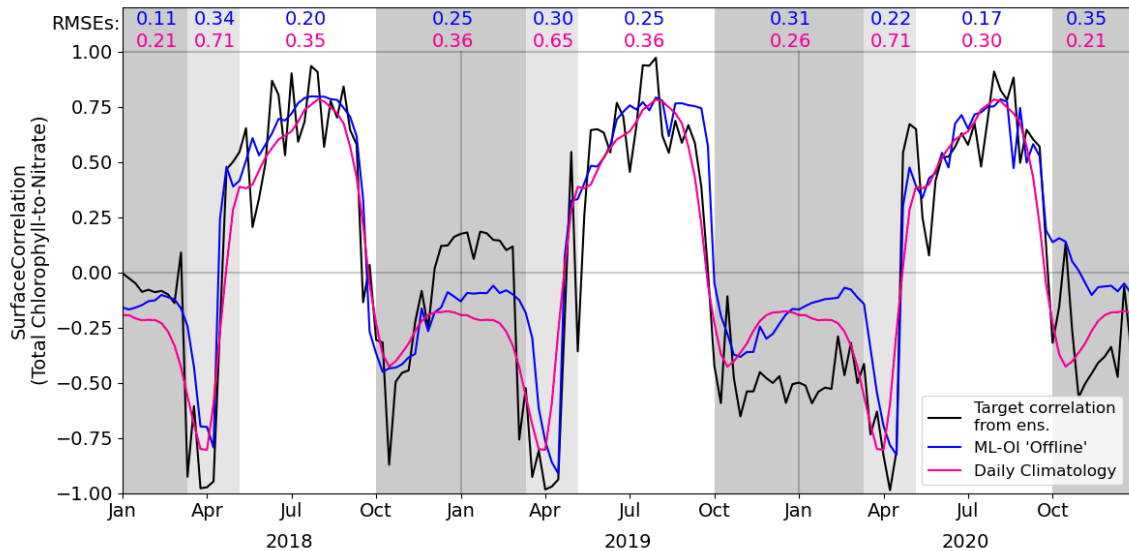


Figure 3: Estimates of correlation between total chlorophyll and nitrate, at weekly intervals across the 3-year offline test period. The target correlation (black) is calculated from the 100-member free-run ensemble (Table 2, row 2). Correlation estimates are shown for ML-OI (blue) and the daily climatology correlations (red; calculated over the 2000–2014 training period). The root mean square errors between the estimated and target correlations are given at the top of the figure, calculated separately over each regime window and for both estimation methods (with corresponding colour). The seasonal regimes of Figure 2 are repeated.

525 The ML-OI model shows clear improvements over climatological estimates of correlation across  
 526 most of the annual cycle. It is important to note, however, that the RMSE here only reflects the  
 527 capability of a method to estimate the ensemble-derived chlorophyll-nitrate correlations. Such im-  
 528 provements do not necessarily translate directly to performance in an online, cycled data assimilation  
 529 system (shown later) where past estimates can influence future states through the model’s dynamical  
 530 evolution.

531 Most clearly, ML-OI is better than climatology for estimating the highly distinctive correlative pattern  
 532 between total chlorophyll and nitrate during the bloom regime, showing a moderate to significant  
 533 RMSE reduction in every bloom period. This pattern consists of a sharp drop to a strongly negative  
 534 correlation, before an almost instantaneous increase to a strong positive correlation. These correlation  
 535 patterns can be simply explained. During the bloom, phytoplankton growth exhausts nutrients, leading  
 536 to negative correlations between chlorophyll and nutrients, whilst the end of the bloom, and the  
 537 following period, phytoplankton growth is nutrient limited, leading to positive correlation. The precise  
 538 timing of the bloom (and hence this correlation pattern) has notable inter-annual variability – varying  
 539 within a period of approximately 5 weeks each year, in this model. The climatological correlations  
 540 estimate this pattern poorly as they are smoothed over this period of inter-annual variability, but the  
 541 ML-OI model, which estimates correlations from the state of the marine BGC model, captures the  
 542 pattern much more accurately.

543 During the nutrient-limited regime, we see a generally strong positive correlation between total  
 544 chlorophyll and nitrate, which has some local variability primarily driven by changes in wind strength,  
 545 such as a weather front passing over the location and mixing nutrients into the surface. The ML-OI  
 546 scheme clearly reduces the RMSE during this period, perhaps capturing some of the local variability  
 547 in this “true” signal and so responds more accurately to these changes in state. Though, it is clear  
 548 that the correlations of the ensemble still vary more strongly than the climatological and ML-based  
 549 correlation estimates.

550 Finally, we also see that during the light-limited regime, the system can exist in either a “weakly  
 551 positive or no correlation” state, or a “moderately negative correlation” state. However, both the  
 552 climatological and ML estimates fail to capture these possible states (both predicting a weakly negative  
 553 correlation over the period). During this time, total chlorophyll and nitrate are generally decoupled,  
 554 and there is no clear link between the state of the system and the correlations estimated. Furthermore,  
 555 as the concentrations of total chlorophyll are very near zero at this time of year (both in the ensemble  
 556 and observations) and there is little spread in the ensemble – the resulting DA updates are small and  
 557 have very little impact. This means that any improvement or degradation in correlation estimates at

558 this time of year are less likely to result in any great improvement to the system, as there is weak  
 559 relationship between chlorophyll and nitrate DA increments and the updates are small. However,  
 560 updates at the start of this period can be important, as the resulting store of nitrate in the upper  
 561 water column could have dynamical impact in later DA cycles when light is no longer limiting and  
 562 the next bloom period starts.

563 After demonstrating the capability of the ML model to estimate the chlorophyll-nitrate relationship  
 564 in an “offline” setting in Fig. 3, in Fig. 4 we compare the performance of a standard EnKF at different  
 565 ensemble sizes with the schemes previously summarised in Sect. 2.4 and 3. This is done in an “online”  
 566 setting, so that any update to the system can have dynamical impact on later DA cycles as the model  
 567 integrates forward in time.

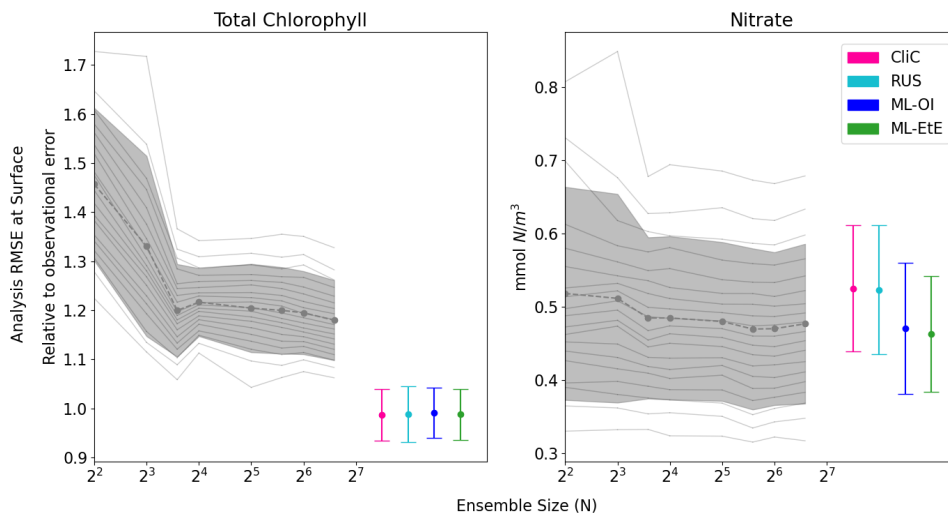


Figure 4: The relationship between analysis RMSE (Eq. (9)) and ensemble size for EnKFs with different ensemble sizes, as well as the performance of the different single-model run schemes. The left panel shows the RMSE of the observed variable, total chlorophyll, normalised relative to the observational error. The right panel shows the RMSE of the unobserved variable, nitrate. The black dashed line represents the mean expected ensemble member error, from an aggregated pool of ensemble members taken from 20 repeat experiments of an EnKF at increasing ensemble sizes, with the shaded grey area indicating  $\pm 1$  standard deviation of ensemble member error. The mean error and  $\pm 1$  standard deviation of 64 independent single-model runs are also given for each of the methods summarised in Sects. 2.4 and 3. An extended version of the plot, showing a wider range of unobserved, not updated variables is given in Fig.B.1.

568 Figure 4 displays the performance of the EnKF, the RUS scheme, the climatological statistics  
 569 scheme (CliC), and the ML schemes. Each panel shows the mean expected error of ensemble members  
 570 for ensemble sizes ranging from 4 to 96. In the left panel for total chlorophyll, the relative RMSE  
 571 is calculated as a ratio of the observation error. The EnKF achieves near-optimal performance for  
 572 ensemble sizes greater than 16, after which the mean expected error reaches a plateau. The relative  
 573 analysis error of total chlorophyll exceeds a value of 1 because it is based on the expected error of  
 574 ensemble members, not the ensemble mean (see Sect. 3.4.1). This reflects the additional stochastic  
 575 error introduced by the EnKF’s generation of perturbed observations. We also see, in the right panel,  
 576 that the error decreases with ensemble size for the unobserved nitrate, indicating that the system  
 577 converges towards more correct nitrate updates at larger ensemble sizes.

578 As expected, the analysis error in the observed total chlorophyll is generally comparable across  
 579 each scheme because they all use the same method, the RUS scheme of Sect. 2.4.1, to update the  
 580 observed total chlorophyll. However, there are more noticeable differences in the schemes that extend  
 581 the updates to nitrate as well. We can clearly see that both the RUS scheme (no update to nitrate)  
 582 and the CliC scheme (update of nitrate using climatological covariances in Eq. (8)) perform similarly  
 583 poorly for nitrate – meaning that the information provided by the observation has not propagated  
 584 well to the unobserved variable. In contrast to this, both ML approaches result in a significant  
 585 improvement in performance, reducing analysis error by between 8 – 12%. This means that the  
 586 information from observations can effectively propagate to the unobserved variables in a single-model  
 587 run, without the need for an expensive ensemble to model the statistics at run time. ML-EtE shows

588 similar improvements during the online testing period as those observed in the offline experiment  
589 (Fig. 3), indicating that the benefits persist even when the updates from each DA cycle feed into  
590 subsequent ones. Moreover, ML-EtE exhibits a lower standard deviation than ML-OI, indicating  
591 reduced sensitivity.

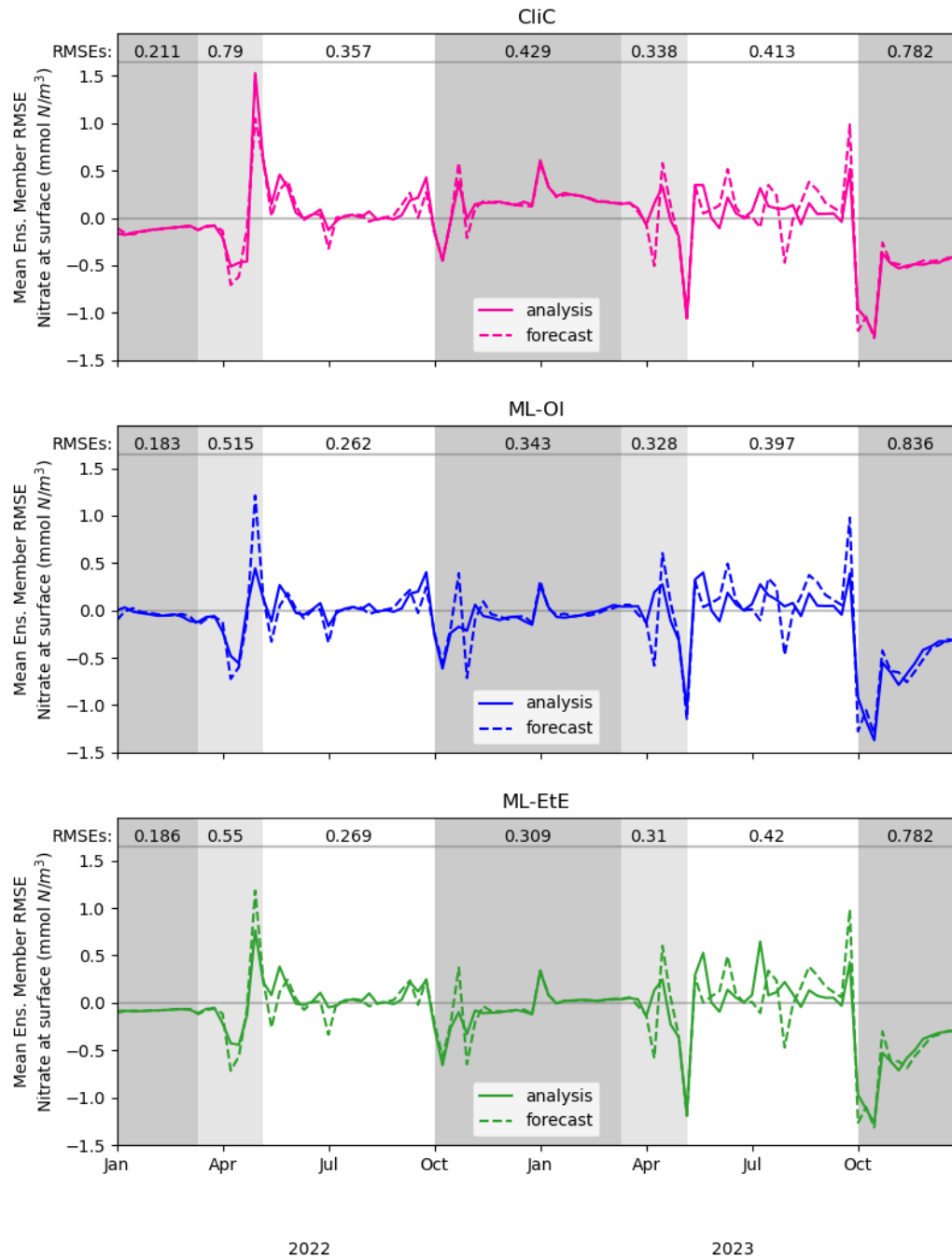


Figure 5: A comparison of mean nitrate background and analysis errors produced in the single-model runs for schemes using “online” cycled-DA at the surface. The first panel shows the analysis RMSE (solid red) and the background RMSE (dashed red) of the CliC runs. The second panel shows the analysis RMSE (solid blue) and the background RMSE (dashed blue) of the ML-OI runs. The third panel shows the analysis RMSE (solid green) and the background RMSE (dashed green) of the ML-EtE runs. For each seasonal regime, the mean analysis RMSE across the entire period and across all initialisations is given at the top of each panel. Shading indicates the system regimes previously outlined in Sect. 4.1: “light-limited” (dark grey), “bloom” (light grey) and “nutrient-limited” (white).

593 ments generated in the “online” setting, and their differences to the truth.

594 While Fig. 4 shows that they improve on average, Fig. 5 gives detail on when improvements are  
 595 made. The runs shown here receive the same observations of the truth, and use the same initial  
 596 conditions and forcing. However, the cycled “online” DA implies that the background state of a  
 597 given time step will differ between methods. Nevertheless, we can see when ML-OI, or ML-EtE,  
 598 make improvements over CliC. A clear example of this is the improved estimations during the bloom  
 599 period, where the ML-estimated methods both provide a lower RMSE than the CliC. This shows that  
 600 both ML methods are able to react to the timing of the bloom event much more accurately than  
 601 climatology can. During the nutrient-limited period, we generally see comparable performance across  
 602 the methods, as the expected correlations are generally high, and the ML-OI method can only weakly  
 603 estimate the variation over this period (as seen previously in Fig. 3). ML-EtE provides no obvious  
 604 advantage in this case, and explains why all methods struggle to make good increments in the second  
 605 year of analyses (where the 7-day forecast errors are typically larger than in the first year). We can  
 606 also see that each approach makes little to no adjustment during the majority of the light-limited  
 607 regime. However, the increments by both ML-OI and ML-EtE at the start of this regime appear to  
 608 yield a prolonged benefit once the system becomes inactive over winter. While these increments are  
 609 unlikely to have any major impact on the system, it is interesting to note that the increments from  
 610 each approach accurately reflect the expected “decoupling” of total chlorophyll to nitrate at this time  
 611 of year. This corroborates with offline experiments of Fig. 3, where the ML-OI model (and indeed,  
 612 the climatological estimates) struggle to replicate the correlations over winter, as there is no strong  
 613 dynamical relationship between the variables at this time.

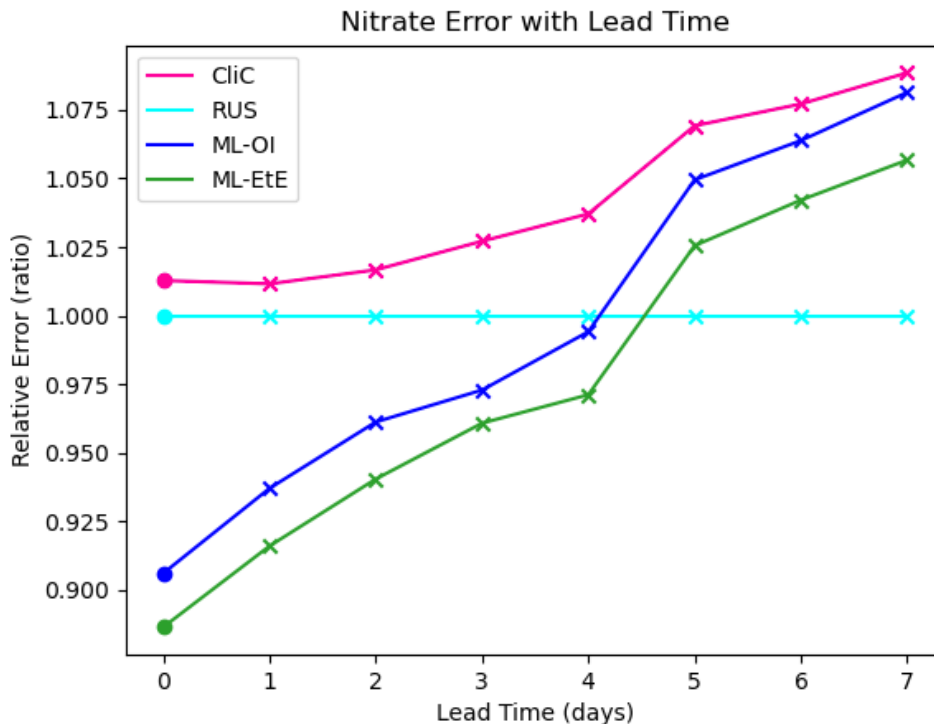


Figure 6: The forecast RMSE of each scheme relative to that of the RUS scheme at daily lead time intervals at the surface. For each scheme, the dot indicates the relative analysis error, while the crosses shows the relative forecast error for each day of lead time until a maximum lead time of 7 days – which is the total time between observations of total chlorophyll in these experiments.

614 Finally, Fig. 6 shows analysis and forecast errors in nitrate in each scheme where errors are nor-  
 615 malised against the error in the RUS scheme. The daily climatological correlations scheme (CliC,  
 616 red line) degrades the analysis error and then makes worse forecasts at every lead time when com-  
 617 pared to the RUS scheme, which does not update the nitrates at all. As previously noted, both ML  
 618 approaches provide an analysis state that is approximately 8-12% better than the not-updated RUS  
 619 nitrate. Improved forecasts then persist for approximately 4-5 days of lead time, only reaching an  
 620 increased relative error after 5 days. For all lead times, ML-EtE outperforms ML-OI. While this is a

621 net benefit to the forecasts of the system, it highlights the difficulty with partially updating a highly  
622 non-linear system. In this, it is clear that each attempt to update the nitrate results in an eventual  
623 error growth beyond simply not updating the system. Part of this could stem from the role of nitrate  
624 as a limiting nutrient; in that it is either available to allow phytoplankton growth, or not. This means  
625 that when estimating an increment for nitrate, we can know that some nitrate should be present or  
626 not, but a precise, continuous quantity that should be added or removed is not information that can  
627 necessarily be inferred from the observation of total chlorophyll. However, this error growth could also  
628 result from the analysis increments introducing some additional imbalance in other quantities of the  
629 system that also need correcting, and the complex marine BGC processes are inter-dependent. These  
630 imbalances and forecast error growths are discussed further in the following Sect. 4.3, when updating  
631 the additional marine BGC variables.

632 In the context of operational systems, such as those implemented by the UK Met Office, total  
633 chlorophyll is assimilated on a daily cycle, and then a forecast is produced for up to six days of lead  
634 time from these improved initial conditions. These results imply that there are huge gains to be  
635 made not only in short term forecasting (before errors saturate again), but also in reanalysis products  
636 that assimilate data with higher frequency, as the ML approaches substantially outperform the RUS  
637 scheme at this point.

### 638 **4.3 Extending the set of updated variables**

639 In this section, we demonstrate the additional benefit of estimating updates not just for nitrate, but  
640 for nearly all marine BGC variables. In Fig. 7, we compare the different ML approaches for updating  
641 an extended set of unobserved marine BGC variables, as well as the previous system that only updates  
642 nitrate.

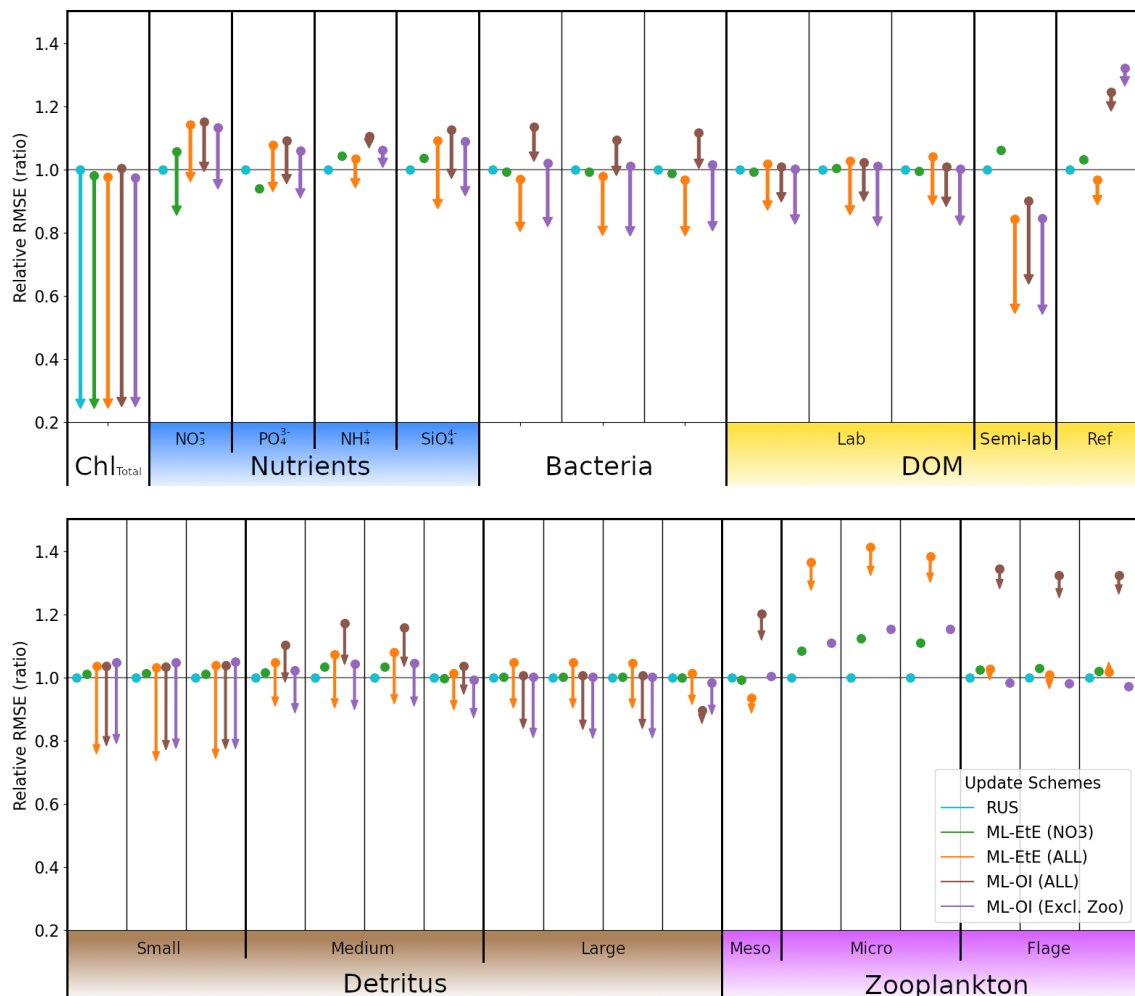


Figure 7: A comparison of the different schemes implemented to update the various components of the ERSEM BGC model at the L4 location. Surface RMSEs are calculated as an average across the entire online test period and across all initialisations. Forecast state (and equivalently background state) RMSEs at 7-day lead times are shown with dots, and the corresponding arrows indicate the analysis RMSEs (no arrow indicates the variable is not updated by the scheme). All RMSEs are relative to the RMSE in the RUS scheme shown in cyan, which only updates the total chlorophyll and PFTs as described in Sect. 2.4.1. The RMSEs of the ML-EtE (NO<sub>3</sub>) scheme, green, are from the same experiment shown previously in Sect. 4.2, and is used as another comparison point for the extended schemes. The ML-EtE (ALL) and ML-OI (ALL), orange and brown respectively, extend the ML schemes described in Sects. 3.2 and 3.1 to update all other pelagic variables. Finally, ML-OI (Excl. Zoo) (purple) updates all pelagic variables, excluding the zooplankton types. The chemical components of each variable class/type follow the same order (left to right) as Table 1.

643 The RUS scheme, described in Sect. 2.4.1, is used as a benchmark for the extended schemes, and  
 644 so values shown in Fig. 7 are RMSEs for 7-day forecasts relative to the RMSE of the RUS method  
 645 (averaged across the 104 forecast-analysis cycles of the entire test period). Again, we recall that the  
 646 RUS does not update any variables beyond total chlorophyll (shown) and its constituent PFTs (not  
 647 shown). The ML-EtE (NO<sub>3</sub>) scheme (green), which updates only nitrate, is carried over from the  
 648 previous section (as it performed best), to act as another point of comparison for the extended schemes.  
 649 Before discussing the extended schemes, we can see from Fig. 7 the dynamical impact that the updates  
 650 of ML-EtE (NO<sub>3</sub>) have on other (i.e. non-updated) marine BGC variables in the system. Generally,  
 651 the change in RMSE for these non-updated variables is very small, with the largest improvement being  
 652 to phosphate and the largest degradation to zooplankton types – particularly microzooplankton. It  
 653 also slightly worsens the forecast error for ammonium and silicate concentrations that are not updated  
 654 during the analysis. While this shows that updating a key nutrient, such as nitrate, can have wider  
 655 impact on the system through dynamical adjustment, the generally beneficial results of the extended  
 656 schemes (discussed below) point towards needing a DA system that can make reasonable adjustments

657 to a wider set of marine BGC variables.

658 Our next scheme, ML-EtE (ALL) (orange), again follows the approach described in Sect. 3.2, but  
659 extends updates to all shown pelagic variables by estimating analysis increments directly from each  
660 background state and total chlorophyll increment. In this L4 setup, this is generally the best perform-  
661 ing scheme, improving unobserved forecast and analysis RMSEs by between 10 – 50%. We emphasise  
662 that while many of the 7-day forecast RMSEs are similar or even degraded compared with RUS, many  
663 of the benefits from an improved analysis persist over a significant portion of the forecast window (as  
664 is shown and discussed later). The most notable exceptions are the zooplankton which, despite having  
665 analysis increments in the correct direction, still return noticeably higher forecast/analysis RMSEs  
666 than most other schemes. Zooplankton have more interactions with other system components, exist-  
667 ing at a higher trophic level, which result in a wider range of uncertainty for their behaviour. This  
668 also suggests they have generally weaker correlations with total chlorophyll. In our configuration, the  
669 RMSEs of the zooplankton group are clearly highly sensitive to updates in other variables, whether  
670 only nitrate is assimilated or a broader set is considered. When correlations between total chlorophyll  
671 and zooplankton are weak (as seen in Fig. A.2), chlorophyll increments contribute little information  
672 for correcting the zooplankton field, which is already strongly influenced by changes elsewhere in the  
673 system.

674 The ML-OI (ALL) scheme (brown) described in 3.1, extends updates to all shown pelagic variables  
675 by estimating the inter-variable correlation from the background state only. This estimation is then  
676 combined with daily varying climatological variances to update the marine BGC state. This method  
677 is also shown to be somewhat effective, generally providing similar behaviour to the ML-EtE (ALL)  
678 scheme, or at least reducing the RMSE from forecast to analysis (even if it is still worse than the  
679 RUS in some cases). It does suffer from the same difficulty in estimating zooplankton updates, to an  
680 even greater degree, which causes some further imbalance in the system. This becomes clear when we  
681 exclude the zooplankton types from the updating in the ML-OI (Excl. Zoo) approach (purple), since  
682 it generally equals or makes small improvements over the ML-EtE (ALL) and ML-OI (ALL) schemes.

683 Figure. A.2 shows that correlations between surface total chlorophyll and silicate (or phosphate) are  
684 as strong as those with nitrate. This suggests that updates of similar magnitude might be expected for  
685 the other nutrients as well. However, it is not straightforward to infer how the system would evolve  
686 when all nutrients are updated simultaneously, based solely on the behaviour observed when only  
687 nitrate is updated. For example, Fig. 7 reveals clear differences in the forecast and analysis RMSE for  
688 nitrate between the two ML-EtE configurations: one updating only nitrate and the other updating  
689 all nutrients. In the case for ML-EtE (ALL), updating all nutrients degrades the nitrate RMSE in  
690 the forecast and provides negligible impact on the analysis, while improving the analysis RMSEs for  
691 phosphate, ammonium and silicate. This outcome is notable, as it may point to assimilation biases  
692 introduced by the choice of update strategy.

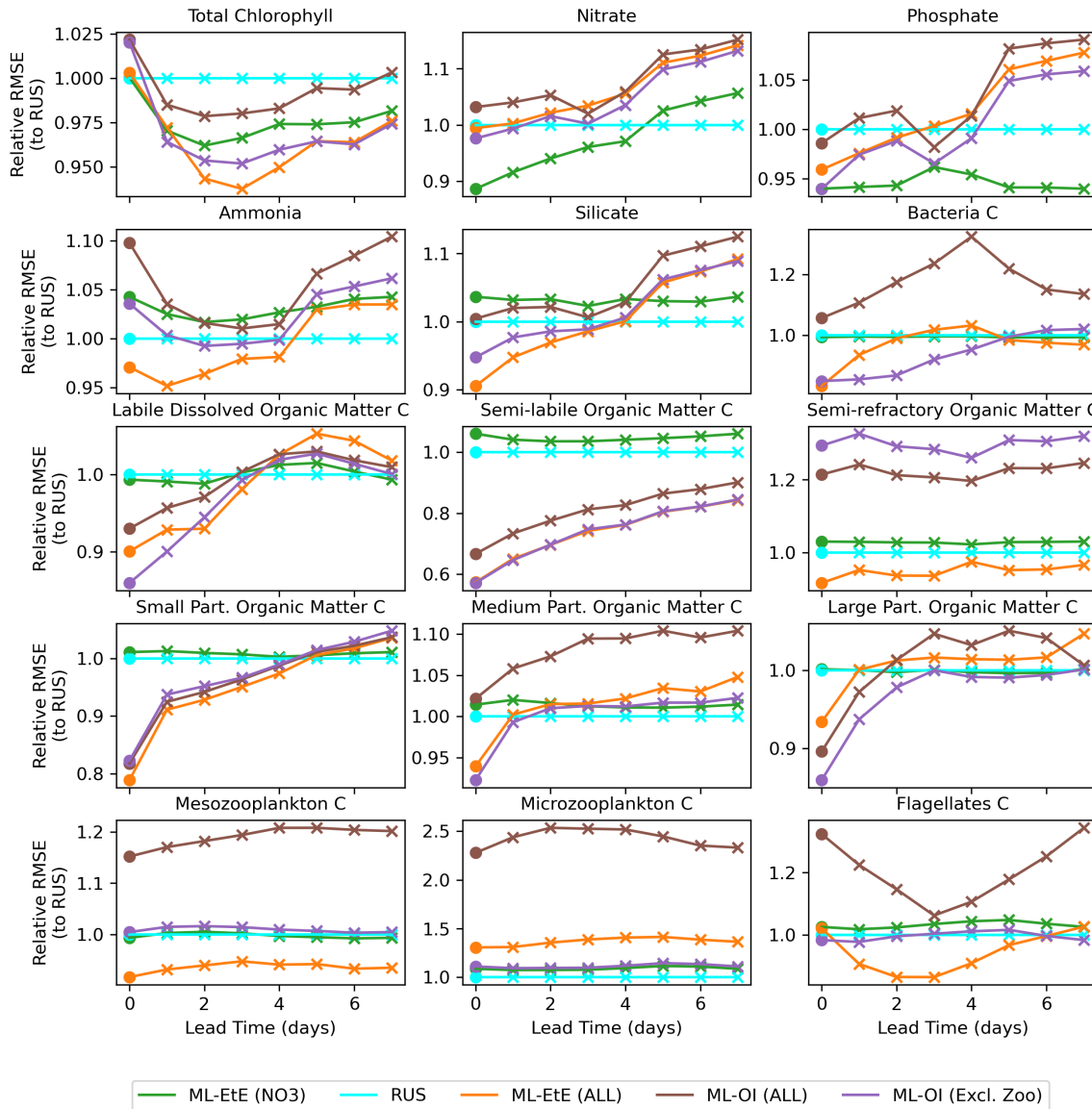


Figure 8: A comparison of the different schemes implemented to update the various components of the ERSEM BGC model at the L4 location (colours used are identical to Fig. 7). The forecast RMSE of each scheme are relative to that of the RUS scheme at daily lead time intervals. The dots indicate the relative analysis error, while the crosses show the relative forecast error for each day of lead time until a maximum lead time of 7 days, which is the time between observations of total chlorophyll.

693 Figure 8 shows the mean RMSE at multiple forecast lead times, for each method relative to the  
 694 RUS scheme. A variety of variables have been selected from Fig. 7 to represent the different behaviours  
 695 observed across each variable group. For many variables, there are notable improvements made over  
 696 the first 3 - 5 forecast lead times (e.g., most schemes for phosphate, ML-EtE (ALL) and ML-OI (Excl.  
 697 Zoo) in silicate, and the particulate organic matter variables), even if the gains do not persist for the  
 698 entire forecast window of 7 days. It is also important to note that the rate at which gains are lost  
 699 is not necessarily quasi-linear for some variables. For example, small particulate organic matter loses  
 700 around half of its 20% gain from the analysis time to 1 day forecast time. Interestingly, the total  
 701 chlorophyll shows similar error relative to the RUS scheme for each other scheme at the analysis time  
 702 (they all handle the observations in the same way, so this should be expected), but each scheme then  
 703 makes improvements at essentially every other forecast lead time. This is likely due to the dynamical  
 704 adjustment of the model (where gains have been made in other variables) propagating through to the  
 705 total chlorophyll values as the model evolves. It is through this dynamical complexity, however, that  
 706 we see other variables do not improve or degrade monotonically as lead time increases (e.g., bacteria  
 707 and flagellates).

708 In summary the experiments from Figs. 7 and 8 show that the impact of DA analysis updates on

709 the model forecast is not straightforward due to the non-linear and complex nature of the BGC model.  
710 Although in general it is true that increasing the number of updated variables benefits the forecasts,  
711 especially at shorter lead times (e.g., less than 5 days), this is definitely not true for every variable.  
712 Some variables, for instance, show similar or worse forecast RMSE at a 7-day lead time than the RUS,  
713 which highlights the need to evaluate how different assimilation strategies and variable-update choices  
714 affect performance.

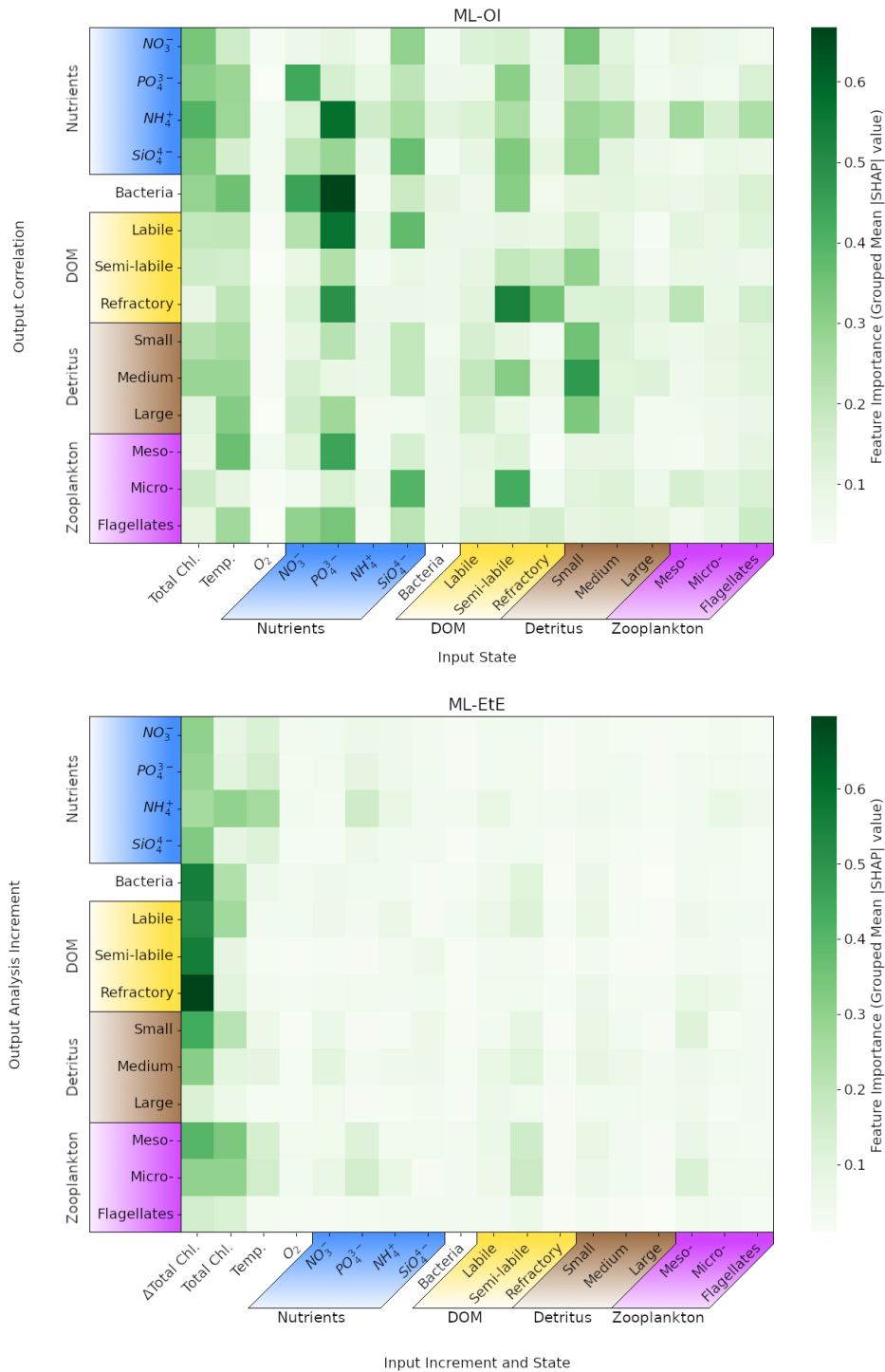


Figure 9: Mean absolute Shapley values estimated for the ML-OI (ALL) and ML-EtE (ALL), across their respective training datasets (see Sect. 3). Chemical components that belong to the same class or type (e.g., the carbon, nitrogen and phosphorus of bacteria) have been grouped as they are highly correlated. The variable names and chemical components are detailed in Table 1. The upper panel shows values for the extended ML-OI (ALL) model, and the lower panel for the extended ML-EtE (ALL) model. The grouped input features for each ML model are given on the  $x$ -axis, which make up the surface state for each pelagic variable. ML-EtE (ALL) has one additional feature,  $\Delta$ Total Chl., which represents the total chlorophyll analysis increment. The grouped output targets for each ML model are given on the  $y$ -axis, which correspond to the correlations between total chlorophyll and each unobserved variable for ML-OI (ALL), and the analysis increments for ML-ETE (ALL).

715 In Fig. 9, we interrogate the ML models using Shapley values (Sect. 3.4.2) to identify important  
 716 ML-model features that are key to making accurate estimations, and drive the connections between

717 observed total chlorophyll and unobserved variables. Such a Shapley analysis also has the potential  
718 to help reduce the number of features needed for training future models (though this was not the aim  
719 of our experiments).

720 Figure 9 shows the grouped mean absolute Shapley values for both the extended ML-OI (upper  
721 panel) and ML-EtE (lower panel) approaches. These are grouped as the separate chemical components  
722 of any class/type, and the resulting Shapley values, are very highly correlated. It is important to note  
723 that Shapley values differ from a pure correlation between the input and output variables. This is  
724 because they capture both direct and interaction effects, account for non-linear relationships, and can  
725 explain a model’s decision-making rather than just measuring statistical association.

726 The ML-OI (ALL) Shapley values indicate that a broad range of input variables are important  
727 to the estimation of total chlorophyll correlations with unobserved variables, and highlight the gen-  
728 eral complexity of these interactions. We see that the state of temperature and total chlorophyll are  
729 moderately important across a broad set of variable groups. This makes sense as this ML model is  
730 estimating the correlation of a given variable with total chlorophyll, which is generally dependent on  
731 the state of total chlorophyll. However, this also implies that the seasonal regimes play a significant  
732 role in the estimations, as temperature is a clear identifier for the current time in the seasonal cycle.  
733 We note that the seasonal signal of other variables could also be important for the estimation of cor-  
734 relations as, in some cases, we see that at least one of the state variables in a group can be important  
735 to estimating the correlation between total chlorophyll and a state variable of the same group. For  
736 example, the state of small detritus is highly important to the entire group of detritus correlations,  
737 the states of some nutrients are generally important to the estimation of nutrients, and the semi-labile  
738 DOM is somewhat important to the wider DOM correlations. Each of the nutrients show moderate  
739 to strong importance across a wide variety of outputs, with the strongest being phosphates. The  
740 elevated Shapley importance of phosphate likely reflects interdependencies among nutrient variables,  
741 particularly given their correlated seasonal behaviour, highlighting a limitation of the Shapley frame-  
742 work when applied to features with shared temporal dynamics. Some input features show no strong  
743 importance to any output targets. In particular, the zooplankton types seem largely unimportant in  
744 estimation their own correlation with total chlorophyll and the variables with a stronger signal have no  
745 obvious direct relationship. This may partially explain why zooplankton performs poorly when they  
746 are updated by the ML-DA schemes, as seen previously in Fig. 7, and points towards the difficulty and  
747 uncertainty associated with zooplankton in marine BGC modelling. This is further evidenced as the  
748 zooplankton types are unimportant as input features for all other correlation estimations as well. We  
749 also see that oxygen is largely unimportant to the estimation of the correlations. This observation is  
750 consistent with the known weak impact of oxygen assimilation in ERSEM on other modelled variables  
751 (Skákala et al., 2021). This would imply that both zooplankton and oxygen could be removed from  
752 the input feature set with little impact on the overall model performance.

753 The ML-EtE (ALL) Shapley values take on a distinctly different structure to those of the ML-OI  
754 (ALL). Recall that ML-EtE (ALL) has a different target than ML-OI (ALL), as it emulates analysis  
755 increments directly. It also has an additional input feature, the analysis increment of total chlorophyll,  
756 which is readily available in both the training dataset and at run-time. The most striking difference  
757 is that the total chlorophyll analysis increment dominates the estimation importances, showing the  
758 highest mean absolute value in almost all estimations. This is to be expected, as the total chlorophyll  
759 increment contains information about the observation, observational error and background model  
760 covariance, which are all necessary components of the unobserved analysis increment as described in  
761 Eq. (1). This makes sense considering the seasonal variation of the model and that total chlorophyll  
762 represents this variation quite reliably according to the regimes discussed in Sect. 4.1. The state  
763 variable input features show much less importance in ML-EtE (ALL) than in ML-OI (ALL), but  
764 they are sometimes still non-zero. These non-zero values seem to correlate somewhat with the most  
765 important input features seen in the ML-OI (ALL) approach, even if they are significantly reduced  
766 overall, suggesting that the state still contributes to the inherent flow dependencies of the analysis  
767 increments.

768 Given the dominant role of the total chlorophyll analysis increment in ML-EtE (ALL), it is impor-  
769 tant to note that the influence of observations could differ when using real observations rather than  
770 synthetic ones, owing to potential differences in error characteristics, representativeness, or systematic  
771 biases. We would still expect the analysis increment of the observed variable to exhibit comparable  
772 importance if trained on increments derived from real observations, as demonstrated in this experi-  
773 ment. However, because of the strong influence of total chlorophyll increments, the overall structure  
774 and nature of the resulting updates may change, reflecting the impact of more complex and realistic  
775 observational error structures.

776 **4.4 Generalisation of machine learned-correlations to an unseen location**

777 In this section, we test the performance of the extended ML approaches from Sect. 4.3 trained for L4,  
 778 but applied to the CWEC location (see Sect. 2.3 and Fig. 1), which exhibits different marine BGC  
 779 behaviour than the L4 training location. In Fig. 10, we assess the performance of these ML models  
 780 according to their 7-day forecast and analysis RMSEs. We then compare some general differences  
 781 between the climatology of the two locations in Fig. 11, and then, with reference to the Shapley  
 782 values shown previously in Fig. 9, we shall discern why the ML model might struggle extrapolating  
 783 to the new location.

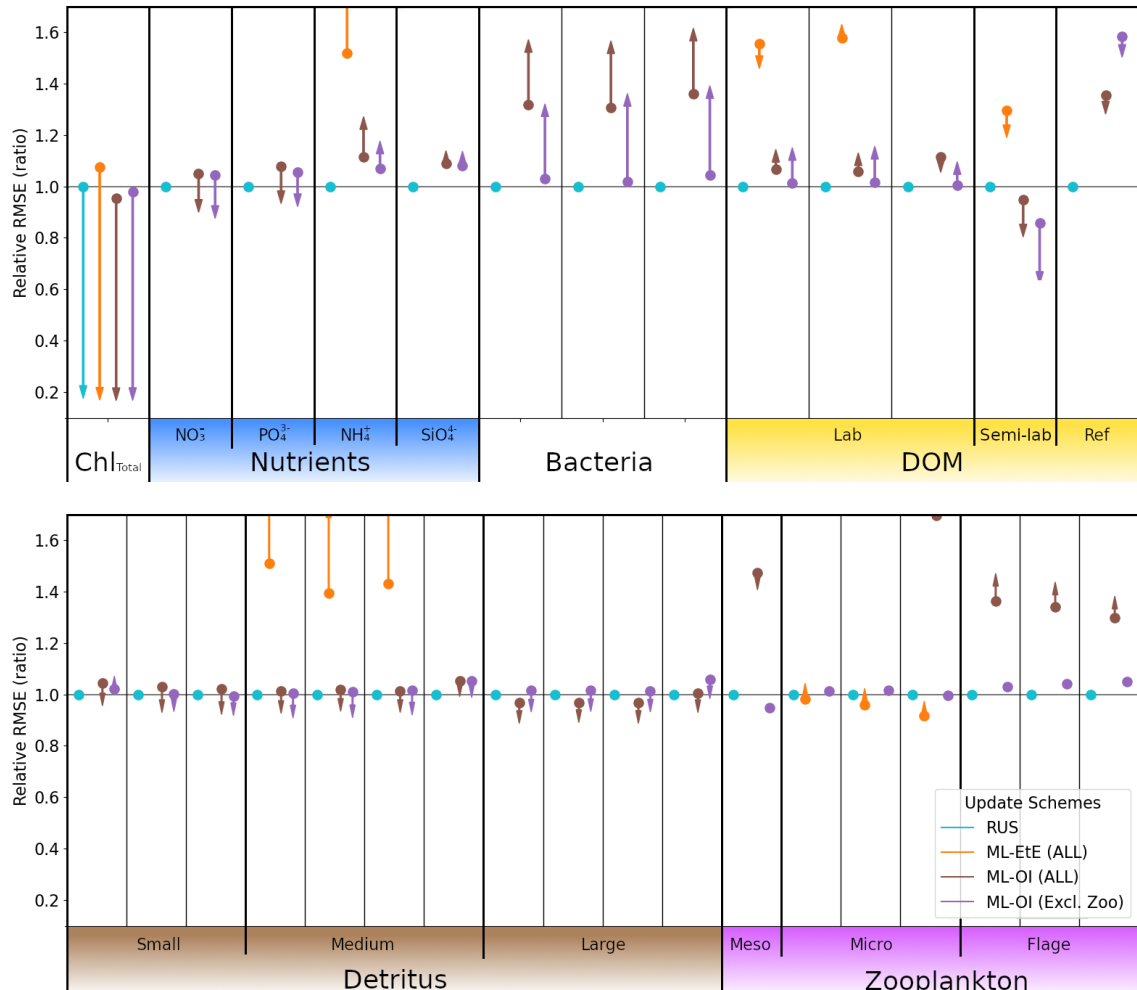


Figure 10: As Fig. 7, with the ML methods trained on L4, but applied to the new location CWEC. Dots and arrows that do not appear are off the scale.

784 Figure 10 again uses the RUS scheme as a comparison point for the ML model approaches, so  
 785 all RMSEs are given as a ratio of the RUS’s RMSE value at the new location. The ML-EtE (ALL)  
 786 approach (orange) performs extremely poorly in this new location, with a large portion of the RMSEs  
 787 exceeding  $1.5\times$  the RUS background error (off the scale of Fig. 10). This is because the emulated  
 788 analysis increments of the EnKF at the L4 location fit the variability and scale of that (trained)  
 789 location and so, do not translate well to the new location. This means that, while the ML-EtE (ALL)  
 790 approach works well at the trained location (and fits the expected distribution of input data), in  
 791 practice its extendability to a new location is limited by both availability of training data and to the  
 792 new location’s similarity to the original training location. The ML-OI (ALL) (brown) makes a marked  
 793 improvement over the ML-EtE (ALL) scheme, which is the reverse of the previous scenario at the  
 794 L4 location. This is likely because the correlations estimated by the ML-OI scheme represent a more  
 795 location-agnostic relationship in the marine BGC variables, which can be used in combination with  
 796 the climatological variances of CWEC to produce more location-appropriate increments (though this  
 797 is not true for ammonium, bacteria and labile DOM, which produce a worse analysis than forecast).  
 798 Furthermore, this scheme still struggles to estimate zooplankton correlations, and so not updating the

799 zooplankton as in the ML-OI (Excl. Zoo) scheme (purple) reduces the damage compared to ML-OI  
800 (ALL) – with any improvements being marginal at best. In both ML-OI (ALL) and ML-OI (Excl.  
801 Zoo), we see that the analysis for detritus is generally improved relative to the RUS scheme (though  
802 these improvements are marginal, and of similar magnitude to the worsened forecasts). Figure 11  
803 shows that the climatological correlations for these variables are generally similar in both locations  
804 (compare Fig. A.2 and Fig. A.3 to see how these vary with time), with small detritus (originating  
805 largely from species with size  $< 20\mu m$ ) showing similarity in both climatological correlation and  
806 standard deviation. Since small detritus is the most important input feature, in Fig. 9, for the  
807 estimation of detritus correlations in the ML-OI models, it is reasonable to see why the improvements  
808 persist between the two locations. We also see in Fig. 10 that both ML-OI models (brown and purple)  
809 make improvements to the analysis RMSEs of nitrate, phosphorus and semi-labile DOM, which show  
810 relatively similar climatological behaviour to L4 in Fig. 11 and Fig. A.1.

811 As all training is performed at one location, it is easy to hypothesise that the ML models are over-  
812 specialised to the characteristics of L4, specifically with regards to their use at other locations. Here,  
813 L4 is coastal and the CWEC is in a more open area of ocean. This does not rule out the possibility  
814 that ML models trained on a limited number of locations could extend their estimations to spatial  
815 locations beyond their set of training locations. However, it indicates that sparse training locations  
816 would need to be chosen carefully, to appropriately cover the spread of behaviour in the system.

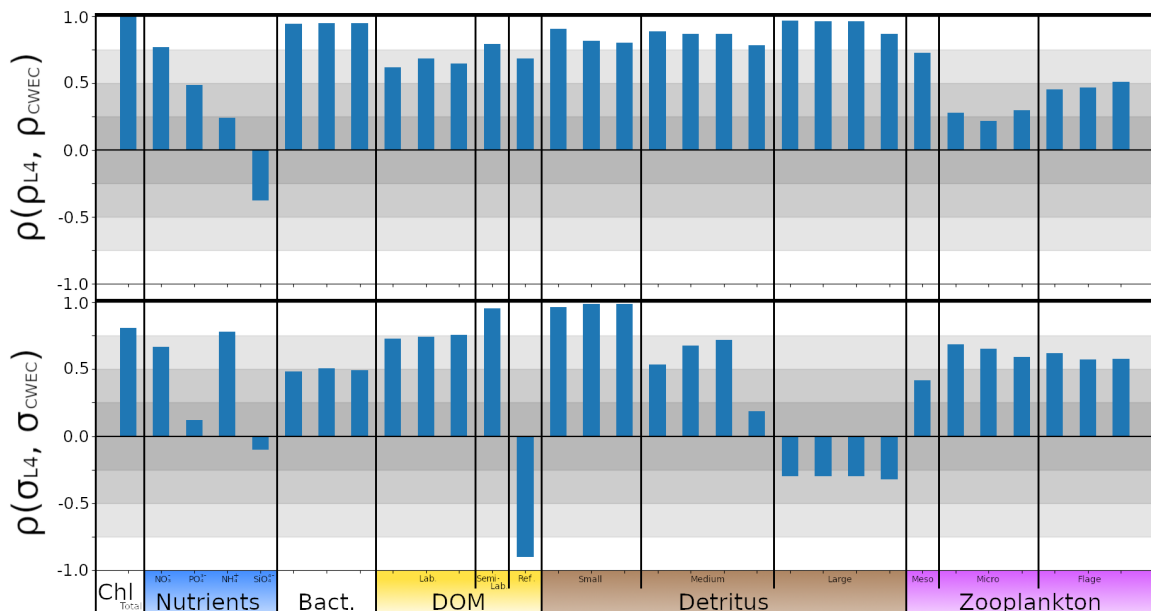


Figure 11: The top panel shows  $\rho(\rho_{L4}, \rho_{CWEC})$ , the correlation between the daily climatological correlations of each variable with total chlorophyll at L4 and at the CWEC. The bottom panel shows the correlation between a given variable's climatological standard deviation signal at L4 and the CWEC,  $\rho(\sigma_{L4}, \sigma_{CWEC})$ . High correlations indicate that the model is behaving similarly in each location.

#### 817 4.5 Viewpoint on scaling multivariate data assimilation to 3D models

818 We have shown that ML methods can make improvements to the DA schemes of marine BGC models  
819 when coupled to a 1D physical model – particularly shown in improved forecasts for shorter lead  
820 times ( $< 4$  days) at the training location. The natural next question is how these results would  
821 scale when the marine BGC model is coupled to a 3D physical model, such as NEMO (Nucleus for  
822 European Modelling of the Ocean). A seemingly simple solution would be to run (once) a well-  
823 tuned, large EnKF, which can then be used to train an ML model to be used operationally in an  
824 analogous 3D DA system that presently updates only total chlorophyll. A reanalysis product with  
825 comprehensive statistics, or all ensemble members available so statistics can be generated, would be  
826 ideal (Bonavita and Laloyaux, 2020; Brajard et al., 2021; Gregory et al., 2024). This circumvents the  
827 need to run an expensive DA scheme operationally as the ML model could be trained offline, and  
828 then run significantly faster while retaining the benefit of the statistics learned from a large ensemble.  
829 This would also allow the analysis increments to be estimated directly. However, state-of-the-art

830 ensemble marine BGC systems are still limited in scale and may not (yet) accurately represent the  
831 statistics needed for multivariate DA (Skákala et al., 2024). Also, this approach would need to be  
832 repeated if/when the observation network changes, which is likely given new observation missions and  
833 strategies (Telszewski et al., 2018). A cheaper alternative would be to calculate the correlations in a  
834 free-run ensemble dataset, as per the methods described in Sect. 3.1. This would be cheaper to create  
835 as there would be no need to store and calculate both background and analysis states. However, this  
836 approach cannot calculate the analysis increments directly and instead must rely on the hybridisation  
837 of background covariances/correlations into existing DA frameworks. Nevertheless our results on the  
838 1D scenario suggests that this is feasible and a good alternative to estimating the increments directly.

839 It is also worth considering how the data for these ML models should be sampled spatially in  
840 the 3D case. In this, we suggest that a sparse forest of 1D models could be generated across the 3D  
841 domain, which aims to cover each region of sufficiently different biogeochemical behaviour. Previous  
842 work by Higgs et al. (2024) has split the North-West European Shelf into dynamically connected  
843 ecoregions, and this, or similar analysis, could be used as a guideline for generating these 1D models.  
844 Furthermore, ML models could handle local multivariate aspects (a 0D transformation), while tradi-  
845 tional DA methods (such as spatial correlations functions) manage 3D reconstruction (just as they  
846 manage the 1D reconstruction in our setup). A limitation of our two test locations is that they are  
847 not directly coupled, and could only be considered weakly coupled in the sense that their forcing data  
848 is extracted from the same 3D weather model. This could mean that 3D models have an advantage  
849 in locations having similar behaviour, as model grid points are much more likely to strongly correlate  
850 due to advection and ocean currents. However, the inverse could also be true, as the 1D models do not  
851 consider riverine input which can have substantial effects at the coast. Either way, the results suggest  
852 that some sparsity could be applied in extracting training data for these models, as long as each  
853 regime of BGC behaviour is represented in the selection. Introducing spatial variables like longitude  
854 and latitude could also improve the models ability to estimate increments or correlations across the  
855 different horizontal locations.

856 An additional avenue worth exploring is whether training ML models on data from multiple,  
857 sufficiently different locations could improve their generalisability. Such an approach would allow for  
858 testing whether a more generalised model can (i) perform as well as a specialised model at its training  
859 locations, and (ii) transfer more successfully to new, unseen locations. While this is beyond the scope  
860 of the present study, it represents an interesting line of future work, particularly when combined  
861 with transfer learning strategies. For instance, a model trained at one location could be used as pre-  
862 conditioning for training at a new site, with the expectation that the pre-conditioned model would  
863 require less additional data to adapt effectively (e.g., Hu et al., 2016). Together, these directions  
864 highlight the importance of balancing specialisation and generalisation when scaling ML-assisted DA  
865 from idealized 1D configurations to realistic 3D systems.

## 866 5 Conclusions

867 Marine biogeochemistry (BGC) models aim to represent the complex BGC processes necessary to  
868 understand and forecast ecosystem behaviour. Data assimilation (DA) plays a crucial role in ensuring  
869 model trajectories remain closely aligned with real-world observations, along with the need for contin-  
870 uous improvement of numerical estimations. In this study, we used a synthetic “perfect model” setup  
871 as a first step to explore ML-assisted DA under controlled conditions, but a natural direction for future  
872 work is to apply these approaches to real observations to assess their practical value. The ML-based  
873 methods considered in this study extend a conventional univariate DA system (RUS, see Table 2),  
874 which observes only total chlorophyll. One ML method (ML-OI) estimates the flow-dependent back-  
875 ground error correlations between total chlorophyll and other model variables (which are then used  
876 in an otherwise conventional DA update step). The other ML method (ML-EtE) directly estimates  
877 the analysis increments of other model variables (bypassing the conventional DA update equations for  
878 these variables).

879 Both numerical modelling and DA are computationally expensive for marine BGC (dealing with  
880 great complexity and many variables), requiring well-tuned and accurately sampled statistics to be  
881 effective. These statistics are often poorly estimated in the undersized ensemble-based methods that  
882 are affordable operationally. In turn, this leads to the use of climatological background error covariance  
883 matrices in deterministic models, or simply not updating unobserved variables. This section concludes  
884 our work in relation to the research questions set out towards the end of Sect. 1 (reproduced below  
885 in italics).

886 *(a) Can we make improvements to the existing univariate scheme by updating a limited set of*

887 *additional variables with an ML model to estimate correlations or analysis increments?* In this study,  
888 we have demonstrated that neural networks can effectively learn statistical relationships between  
889 total chlorophyll (the only observed variable) and various pelagic BGC model variables. With ma-  
890 chine learning (ML), we achieve improvements over climatological statistics in the nitrate-only update  
891 framework. Our analysis of ML-estimated nitrate updates illustrates that the ML methods behave in  
892 a largely coherent and meaningful manner. While ML can degrade forecast skill in some unobserved  
893 variables compared to RUS or nitrate-only ML schemes, they nonetheless show promise, with their  
894 usefulness in more complex assimilation settings requiring further assessment.

895 *(b) Can these ML models be extended to effectively update all unobserved pelagic variables?* ML  
896 models can update almost all unobserved pelagic variables, supporting the broader applicability of  
897 ML in DA. In our configuration, zooplankton does not update well using either of the ML methods  
898 extended to all state variables (ML-OI (ALL) and ML-EtE (ALL)), and is better treated in hybrid  
899 DA schemes without being updated directly (as in ML-OI (Excl. Zoo)). This limitation may reflect  
900 the particular parameterisations of zooplankton–phytoplankton interactions, grazing, and mortality  
901 in the underlying BGC model, and may not generalise to other model setups. More broadly, we expect  
902 that variables less directly linked to or less sensitive to the observed quantity will be more difficult  
903 to update well. Since parameterisation choices can alter these relationships, new parameterisations  
904 would likely require retraining emulators, or alternatively, more flexible ML strategies such as transfer  
905 learning. Exploring such approaches in more diverse configurations remains an important avenue for  
906 follow-on investigations.

907 *(c) Is the ML model transferable to a new location after being trained on some other location?*  
908 While a neural network trained in one water column exhibits partial transferability to other locations,  
909 challenges remain in fully generalising the model across spatial domains. This partial transferability  
910 is valuable, given the difficulty and cost of acquiring high-quality training data across large oceanic  
911 regions, and should be explored further in the context of 3D models. We discuss the feasibility of  
912 this, and propose a methodology for doing so. Future work should focus on refining transferability  
913 strategies, effective sampling strategy to allow for ergodic coverage (i.e., ensuring statistical represen-  
914 tativeness over space and/or time), and further evaluating the scalability of ML-driven DA in complex  
915 marine environments.

916 **A Characterisation of location biogeochemistry**

917 Figure A.1 shows the variability of each ERSEM variable over the online testing period for the L4  
 918 and CWEC locations. It shows that the CWEC is clearly less biologically productive than L4 with  
 919 surface concentrations of total chlorophyll having a significantly lower median value, and a maximum  
 920 that is approximately 50% of L4’s maximum. Each exhibits similar temperature values, as they are  
 921 both located within the English Channel. In the nutrients group, nitrate and phosphate values cover  
 922 a similar range in each location, but ammonium and silicate have little overlap. Bacteria and DOM  
 923 concentrations also show little similarity between locations. The small detritus concentrations are very  
 924 similar between both locations, but the medium and large detritus differ significantly, with CWEC  
 925 covering a much wider range of values than L4. Zooplankton concentrations also differ between the  
 926 locations, with CWEC producing much lower concentrations of zooplankton than the more biologically  
 927 active L4 location.

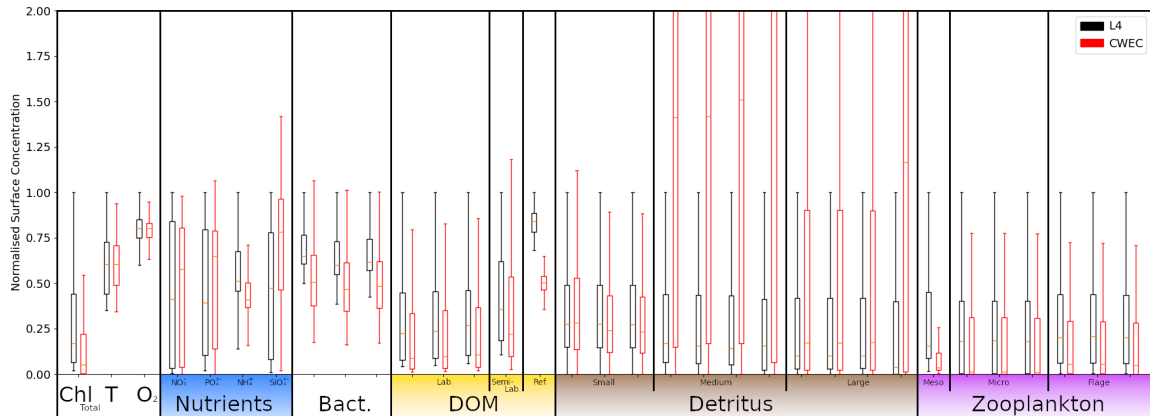


Figure A.1: Box and whisker plot showing the 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentile and upper and lower bound (excluding outliers larger than 1.5 × Inter quartile range) of each pelagic marine BGC variable for the RUS scheme in the online testing period. Values for L4 are given in black, and values for the CWEC are given in red. All values are normalised against the upper bound of L4. Chemical components are ordered according to Table 1. The label “T” corresponds to temperature.

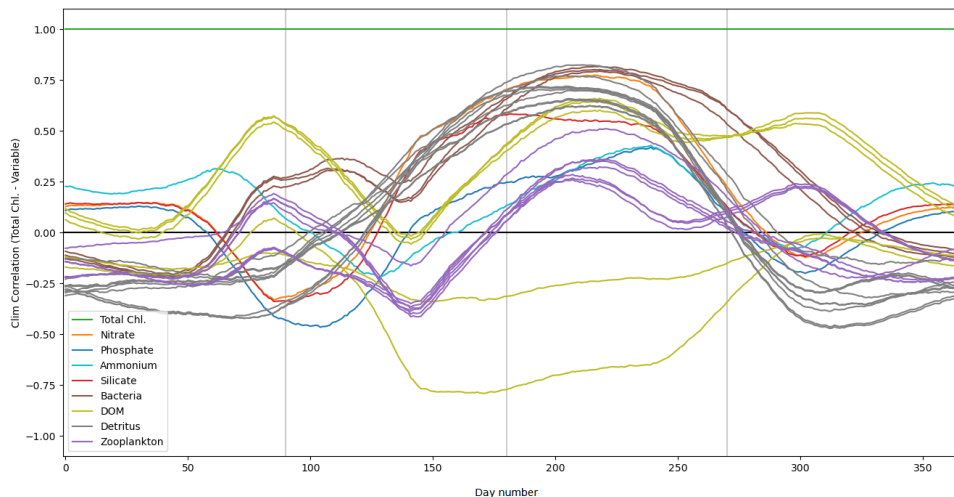


Figure A.2: Daily climatological correlations between total chlorophyll and each pelagic variable at the L4 location, calculated from the training free-run period of 2000-2014. Pelagic variables of the same class (according to Table 2) are shown with the same colour, except nutrients (nitrate, phosphate, ammonium and silicate) which are shown with separate colours.

928 The climatological correlations between total chlorophyll and other pelagic variables at the L4

929 location, shown in Fig. A.2, vary significantly according to the season. Variables of the same class  
 930 (see Table 1) generally exhibit very similar correlations. Correlations are much stronger during the  
 931 spring and summer months, as this period is more biologically active, and so the different model  
 932 components are going to be more closely coupled. Some variables, such as zooplankton, show a much  
 933 weaker correlative relationship with total chlorophyll.

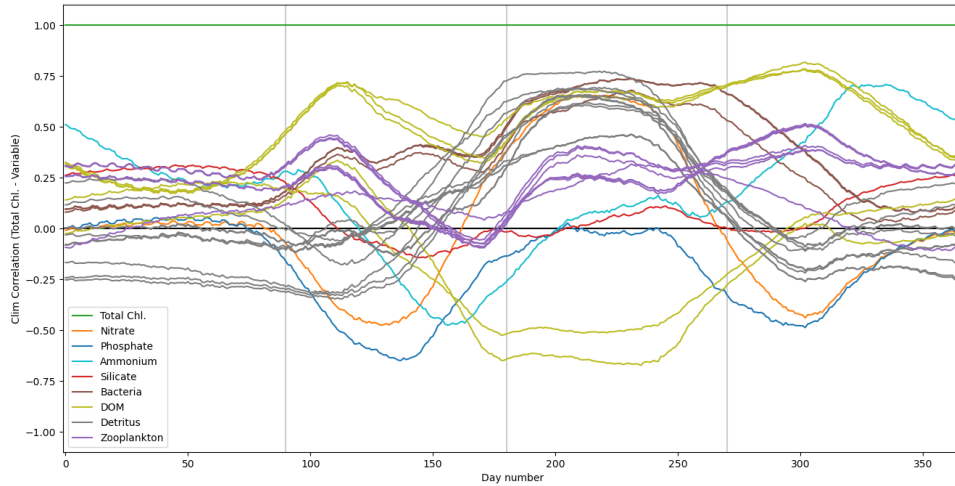


Figure A.3: As with Fig.A.2, except for the CWEC location from a free-run period of 2000-2010.

934 Figure A.3 shows the climatological correlations between total chlorophyll and other pelagic vari-  
 935 ables at the CWEC location. As with L4, the correlations of most variables show a much stronger  
 936 correlation with total chlorophyll during the spring and summer, when the system is much more ac-  
 937 tive. The correlations of nitrate are similar to those seen at the L4 location in Fig. A.2, following the  
 938 pattern described Sect. 4.1. Zooplankton shows a weak correlation with total chlorophyll.

## 939 B Dynamical impact of updating only nitrate

940 Figure B.1 is an extension of Fig. 4, with a representative set of variables that are unobserved and not  
 941 updated (unlike nitrate which is also unobserved, but updated in this experiment). This clearly shows  
 942 that the improvement of nitrate does not necessarily translate to an improvement in other variables,  
 943 regardless of the method used to update the nitrate. This highlights the need to evaluate how different  
 944 assimilation strategies and variable-update choices affect performance.

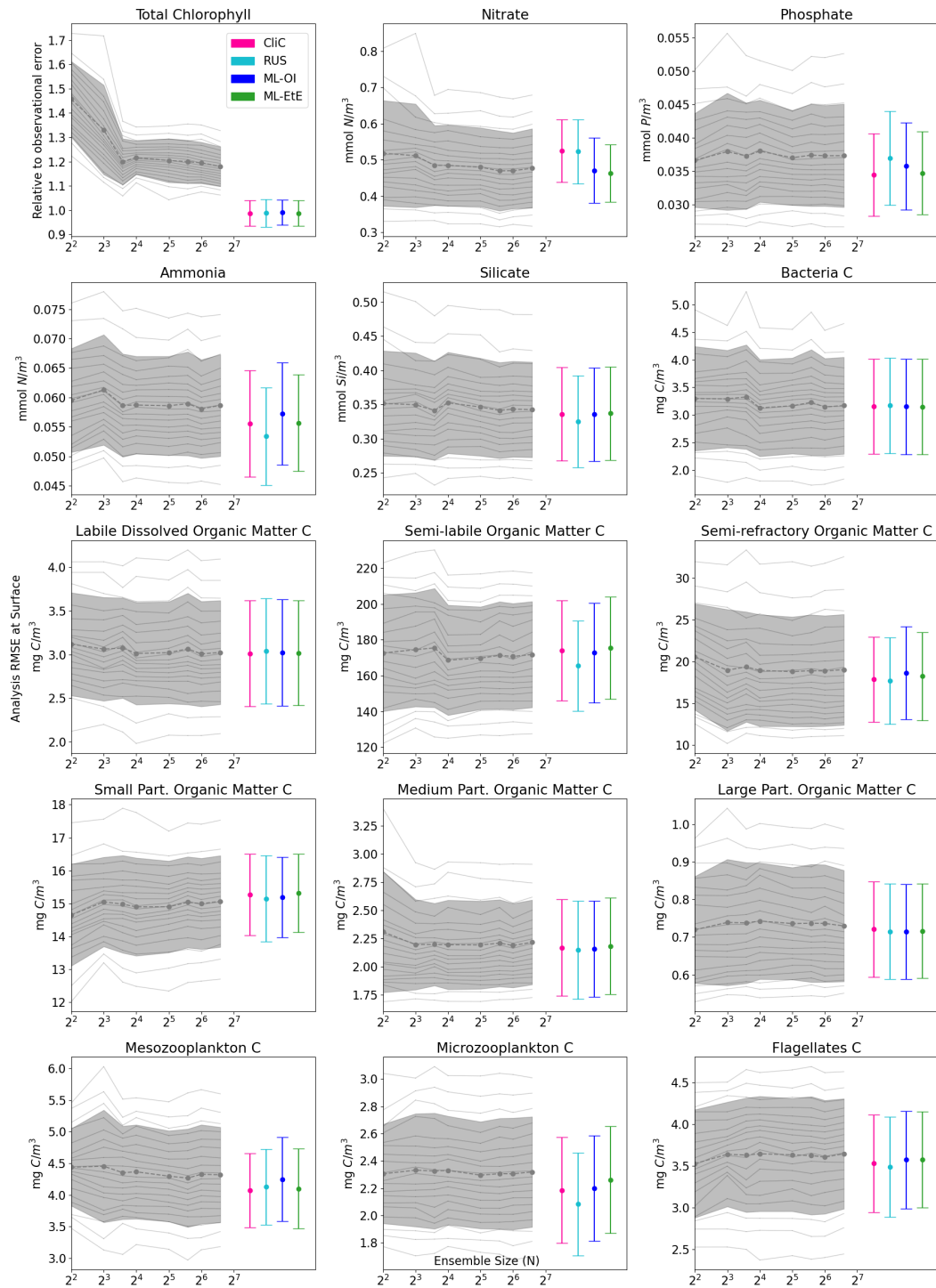


Figure B.1: An extension of Fig. 4, where panels for total chlorophyll (observed) and nitrate (unobserved, but updated) are the same. representative set of additional variables (unobserved, not updated) are shown. A “C” indicates the panel is given for the carbon chemical component in the model.

## 945 Author Contributions

946 IH wrote and executed all code. All authors contributed to analysing and interpreting the results,  
 947 proof-reading the manuscript, and adjusting the text.

## <sup>948</sup> **Competing interests**

<sup>949</sup> One of the (co-)authors is a member of the editorial board of Biogeosciences. The peer-review process  
<sup>950</sup> was guided by an independent editor, and the authors also have no other competing interests to  
<sup>951</sup> declare.

## References

- 952
- 953 Anugerahanti, P., Kerimoglu, O., and Smith, S. L. (2021). Enhancing ocean biogeochemical models  
954 with phytoplankton variable composition. *Frontiers in Marine Science*, 8:944.
- 955 Artioli, Y., Blackford, J. C., Butenschön, M., Holt, J. T., Wakelin, S. L., Thomas, H., Borges, A. V.,  
956 and Allen, J. I. (2012). The carbonate system in the North Sea: Sensitivity and model validation.  
957 *Journal of Marine Systems*, 102:1–13.
- 958 Asch, M., Bocquet, M., and Nodet, M. (2016). *Data assimilation: methods, algorithms, and applica-*  
959 *tions*. SIAM.
- 960 Baretta, J., Ebenhöf, W., and Ruardij, P. (1995). The European regional seas ecosystem model, a  
961 complex marine ecosystem model. *Netherlands Journal of Sea Research*, 33(3-4):233–246.
- 962 Baretta-Bekker, J., Baretta, J., and Ebenhöf, W. (1997). Microbial dynamics in the marine ecosystem  
963 model ERSEM II with decoupled carbon assimilation and nutrient uptake. *Journal of Sea Research*,  
964 38(3-4):195–211.
- 965 Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M. (2020). DINCAE 1.0: A convolutional  
966 neural network with error estimates to reconstruct sea surface temperature satellite observations.  
967 *Geoscientific Model Development*, 13(3):1609–1622.
- 968 Bertino, L., Ali, A., Carrasco, A., Lien, V., and Melsom, A. (2021). The Arctic Marine Forecasting  
969 Center in the first Copernicus period. In *9th EuroGOOS International conference*, pages 256–263.
- 970 Blackford, J. (1997). An analysis of benthic biological dynamics in a North Sea ecosystem model.  
971 *Journal of Sea Research*, 38(3-4):213–230.
- 972 Bocquet, M., Farchi, A., Finn, T. S., Durand, C., Cheng, S., Chen, Y., Pasmans, I., and Carrassi,  
973 A. (2024). Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities  
974 without an ensemble. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(9).
- 975 Bolding, K. and Villarreal, M. R. (1999). GOTM: A general ocean turbulence model: Theory, appli-  
976 cations and test cases. Technical report, European Commission Tech. Rep. EUR 18745 EN.
- 977 Bonavita, M. and Laloyaux, P. (2020). Machine learning for model error inference and correction.  
978 *Journal of Advances in Modeling Earth Systems*, 12(12):e2020MS002232.
- 979 Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2020). Combining data assimilation and  
980 machine learning to emulate a dynamical model from sparse and noisy observations: A case study  
981 with the Lorenz 96 model. *Journal of computational science*, 44:101171.
- 982 Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2021). Combining data assimilation and  
983 machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal*  
984 *Society A*, 379(2194):20200086.
- 985 Bruggeman, J. and Bolding, K. (2014). A general framework for aquatic biogeochemical models.  
986 *Environmental modelling & software*, 61:249–265.
- 987 Bruggeman, J., Bolding, K., Nerger, L., Teruzzi, A., Spada, S., Skakala, J., Ciavatta, S., et al.  
988 (2024). Eat v1. 0.0: a 1d test bed for physical–biogeochemical data assimilation in natural waters.  
989 *GEOSCIENTIFIC MODEL DEVELOPMENT*, 17(14):5619–5639.
- 990 Buizza, C., Casas, C. Q., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur,  
991 T., Ruiz, L. B., et al. (2022). Data learning: Integrating data assimilation and machine learning.  
992 *Journal of Computational Science*, 58:101525.
- 993 Burchard, H. (2009). Combined effects of wind, tide, and horizontal density gradients on stratification  
994 in estuaries and coastal seas. *Journal of Physical Oceanography*, 39(9):2117–2136.
- 995 Butenschön, M., Clark, J., Aldridge, J. N., Allen, J. I., Artioli, Y., Blackford, J., Bruggeman, J.,  
996 Cazenave, P., Ciavatta, S., Kay, S., et al. (2016). ERSEM 15.06: a generic model for marine biogeo-  
997 chemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development*,  
998 9(4):1293–1339.

- 999 Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences:  
1000 An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*,  
1001 9(5):e535.
- 1002 Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss,  
1003 B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., and Arcucci,  
1004 R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical  
1005 systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387.
- 1006 Ciavatta, S., Brewin, R., Skakala, J., Polimene, L., de Mora, L., Artioli, Y., and Allen, J. I. (2018).  
1007 Assimilation of ocean-color plankton functional types to improve marine ecosystem simulations.  
1008 *Journal of Geophysical Research: Oceans*, 123(2):834–854.
- 1009 Ciavatta, S., Kay, S., Brewin, R. J., Cox, R., Di Cicco, A., Nencioli, F., Polimene, L., Sammartino, M.,  
1010 Santoleri, R., Skakala, J., et al. (2019). Ecoregions in the Mediterranean Sea through the reanalysis  
1011 of phytoplankton functional types and carbon fluxes. *Journal of Geophysical Research: Oceans*,  
1012 124(10):6737–6759.
- 1013 Ciavatta, S., Kay, S., Saux-Picart, S., Butenschön, M., and Allen, J. (2016). Decadal reanalysis of  
1014 biogeochemical indicators and fluxes in the North West European shelf-sea ecosystem. *Journal of*  
1015 *Geophysical Research: Oceans*, 121(3):1824–1845.
- 1016 Ciavatta, S., Torres, R., Martinez-Vicente, V., Smyth, T., Dall’Olmo, G., Polimene, L., and Allen,  
1017 J. I. (2014). Assimilation of remotely-sensed optical properties to improve marine biogeochemistry  
1018 modelling. *Progress in Oceanography*, 127:74–95.
- 1019 Ciliberti, S. A., Grégoire, M., Staneva, J., Palazov, A., Coppini, G., Lecci, R., Peneva, E., Matreata,  
1020 M., Marinova, V., Masina, S., et al. (2021). Monitoring and forecasting the ocean state and bio-  
1021 geochemical processes in the Black Sea: Recent developments in the Copernicus Marine Service.  
1022 *Journal of Marine Science and Engineering*, 9(10):1146.
- 1023 Coppini, G., Clementi, E., Cossarini, G., Korres, G., Drudi, M., Amadio, C., Aydogdu, A., Agostini,  
1024 P., Bolzon, G., Cretì, S., et al. (2021). The Copernicus marine service ocean forecasting system for  
1025 the Mediterranean Sea. In *9th EuroGOOS International conference*, pages 272–279.
- 1026 Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., and  
1027 d’Ortenzio, F. (2019). Towards operational 3D-Var assimilation of chlorophyll biogeochemical-Argo  
1028 float data into a biogeochemical model of the Mediterranean Sea. *Ocean Modelling*, 133:112–128.
- 1029 Cossarini, G., Querin, S., Solidoro, C., Sannino, G., Lazzari, P., Di Biagio, V., and Bolzon, G. (2017).  
1030 Development of BFMCOUPLER (v1. 0), the coupling scheme that links the MITgcm and BFM  
1031 models for ocean biogeochemistry simulations. *Geoscientific Model Development*, 10(4):1423–1445.
- 1032 Council, N. R., on Geosciences, C., Science, W., Board, T., Board, O. S., on the Causes, C., and  
1033 of Coastal Eutrophication, M. (2000). *Clean Coastal Waters: Understanding and Reducing the*  
1034 *Effects of Nutrient Pollution*. National Academies Press.
- 1035 Doney, S. C., Fabry, V. J., Feely, R. A., and Kleypas, J. A. (2009). Ocean acidification: the other  
1036 CO<sub>2</sub> problem. *Annual review of marine science*, 1(1):169–192.
- 1037 Dowd, M., Jones, E., and Parslow, J. (2014). A statistical overview and perspectives on data assimi-  
1038 lation for marine biogeochemical models. *Environmetrics*, 25(4):203–213.
- 1039 Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementa-  
1040 tion. *Ocean dynamics*, 53:343–367.
- 1041 Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., and Rousseau, F. (2021). Learning  
1042 variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*,  
1043 13(10):e2021MS002572.
- 1044 Falchetti, S., Conley, D. C., Brocchini, M., and Elgar, S. (2010). Nearshore bar migration and  
1045 sediment-induced buoyancy effects. *Continental Shelf Research*, 30(2):226–238.
- 1046 Fennel, K., Gehlen, M., Brasseur, P., Brown, C. W., Ciavatta, S., Cossarini, G., Crise, A., Edwards,  
1047 C. A., Ford, D., Friedrichs, M. A., et al. (2019). Advancing marine biogeochemical and ecosystem  
1048 reanalyses and forecasts as tools for monitoring and managing ecosystem health. *Frontiers in Marine*  
1049 *Science*, 6:89.

- 1050 Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., and Yu, L. (2022). Ocean  
1051 biogeochemical modelling. *Nature Reviews Methods Primers*, 2(1):76.
- 1052 Fennel, K. and Testa, J. M. (2019). Biogeochemical controls on coastal hypoxia. *Annual Review of*  
1053 *Marine Science*, 11(1):105–130.
- 1054 Ford, D., Edwards, K., Lea, D., Barciela, R., Martin, M., and Demaria, J. (2012). Assimilating  
1055 GlobColour ocean colour data into a pre-operational physical-biogeochemical model. *Ocean Science*,  
1056 8(5):751–771.
- 1057 Ford, D., Key, S., McEwan, R., Totterdell, I., and Gehlen, M. (2018). Marine biogeochemical modelling  
1058 and data assimilation for operational forecasting, reanalysis, and climate research. *New Frontiers*  
1059 *in Operational Oceanography*, pages 625–652.
- 1060 Frölicher, T. L. and Laufkötter, C. (2018). Emerging risks from marine heat waves. *Nature commu-*  
1061 *nications*, 9(1):650.
- 1062 Galli, G., Wakelin, S., Harle, J., Holt, J., and Artioli, Y. (2024). Multi-model comparison of trends  
1063 and controls of near-bed oxygen concentration on the Northwest European Continental Shelf under  
1064 climate change. *Biogeosciences*, 21(8):2143–2158.
- 1065 Gehlen, M., Barciela, R., Bertino, L., Bresseur, P., Butenschön, M., Chai, F., Crise, A., Drillet, Y.,  
1066 Ford, D., Lavoie, D., et al. (2015). Building the capacity for forecasting marine biogeochemistry  
1067 and ecosystems: recent advances and future developments. *Journal of Operational Oceanography*,  
1068 8(sup1):s168–s187.
- 1069 Geider, R., MacIntyre, H., and Kana, T. (1997). Dynamic model of phytoplankton growth and  
1070 acclimation: responses of the balanced growth rate and the chlorophyll a: carbon ratio to light,  
1071 nutrient-limitation and temperature. *Marine Ecology Progress Series*, 148:187–200.
- 1072 Gobler, C. J. (2020). Climate change and harmful algal blooms: insights and perspective. *Harmful*  
1073 *algae*, 91:101731.
- 1074 Gregg, W. W. and Rousseaux, C. S. (2017). Simulating pace global ocean radiances. *Frontiers in*  
1075 *Marine Science*, 4:60.
- 1076 Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., and Zanna, L. (2024). Machine learning for  
1077 online sea ice bias correction within global ice-ocean simulations. *Geophysical Research Letters*,  
1078 51(3):e2023GL106776.
- 1079 Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., Brockmann, C.,  
1080 Chauhan, P., Choi, J.-k., Chuprin, A., et al. (2019). Satellite ocean colour: Current status and  
1081 future perspective. *Frontiers in Marine Science*, 6:485.
- 1082 Gutknecht, E., Refray, G., Mignot, A., Dabrowski, T., and Sotillo, M. G. (2019). Modelling the marine  
1083 ecosystem of Iberia–Biscay–Ireland (IBI) European waters for CMEMS operational applications.  
1084 *Ocean Science*, 15(6):1489–1516.
- 1085 Hayashida, H., Steiner, N., Monahan, A., Galindo, V., Lizotte, M., and Lavoie, M. (2017). Impli-  
1086 cations of sea-ice biogeochemistry for oceanic production and emissions of dimethyl sulfide in the  
1087 Arctic. *Biogeosciences*, 14(12):3129–3155.
- 1088 Heinze, C. and Gehlen, M. (2013). Modeling ocean biogeochemical processes and the resulting tracer  
1089 distributions. In *International Geophysics*, volume 103, pages 667–694. Elsevier.
- 1090 Hemmings, J. C., Barciela, R. M., and Bell, M. J. (2008). Ocean color data assimilation with material  
1091 conservation for improving model estimates of air-sea CO<sub>2</sub> flux.
- 1092 Higgs, I., Skákala, J., Bannister, R., Carrassi, A., and Ciavatta, S. (2024). Investigating ecosystem  
1093 connections in the shelf sea environment using complex networks. *Biogeosciences*, 21(3):731–746.
- 1094 Hu, Q., Zhang, R., and Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with  
1095 deep neural networks. *Renewable Energy*, 85:83–95.
- 1096 Jin, H., Song, Q., and Hu, X. (2019). Auto-keras: An efficient neural architecture search system.  
1097 In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data*  
1098 *mining*, pages 1946–1956.

- 1099 Jones, E. M., Baird, M. E., Mongin, M., Parslow, J., Skerratt, J., Lovell, J., Margvelashvili, N.,  
1100 Matear, R. J., Wild-Allen, K., Robson, B., et al. (2016). Use of remote-sensing reflectance to con-  
1101 strain a data assimilating marine biogeochemical model of the Great Barrier Reef. *Biogeosciences*,  
1102 13(23):6441–6469.
- 1103 Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects.  
1104 *Science*, 349(6245):255–260.
- 1105 Kerimoglu, O., Anugerahanti, P., and Smith, S. L. (2021). FABM-NflexPD 1.0: assessing an instantane-  
1106 ous acclimation approach for modeling phytoplankton growth. *Geoscientific Model Development*,  
1107 14(10):6025–6047.
- 1108 Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., and Hoyer, S. (2021). Machine  
1109 learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sci-  
1110 ences*, 118(21):e2101784118.
- 1111 Le Traon, P. Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., Bentamy,  
1112 A., Bertino, L., Brando, V. E., Kreiner, M. B., et al. (2019). From observation to information and  
1113 users: The copernicus marine service perspective. *Frontiers in Marine Science*, 6:234.
- 1114 Leeds, W., Wikle, C., Fiechter, J., Brown, J., and Milliff, R. (2013). Modeling 3-D spatio-temporal  
1115 biogeochemical processes with a forest of 1-D statistical emulators. *Environmetrics*, 24(1):1–12.
- 1116 Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances  
1117 in neural information processing systems*, 30.
- 1118 Mandal, S., Homma, H., Priyadarshi, A., Burchard, H., Smith, S. L., Wirtz, K. W., and Yamazaki,  
1119 H. (2016). A 1D physical–biological model of the impact of highly intermittent phytoplankton  
1120 distributions. *Journal of Plankton Research*, 38(4):964–976.
- 1121 Mattern, J. P., Fennel, K., and Dowd, M. (2012). Estimating time-dependent parameters for a  
1122 biological ocean model using an emulator approach. *Journal of Marine Systems*, 96:32–47.
- 1123 Mattern, J. P., Fennel, K., and Dowd, M. (2013). Sensitivity and uncertainty analysis of model hypoxia  
1124 estimates for the Texas-Louisiana shelf. *Journal of Geophysical Research: Oceans*, 118(3):1316–  
1125 1332.
- 1126 Mattern, J. P., Fennel, K., and Dowd, M. (2014). Periodic time-dependent parameters improving  
1127 forecasting abilities of biological ocean models. *Geophysical Research Letters*, 41(19):6848–6854.
- 1128 Mattern, J. P., Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J. (2017). Data assimilation of  
1129 physical and chlorophyll a observations in the California Current System using two biogeochemical  
1130 models. *Ocean Modelling*, 109:55–71.
- 1131 McEwan, R., Kay, S., and Ford, D. (2021). Quality information document for the CMEMS North  
1132 West European Shelf biogeochemical analysis and forecast. Technical report, CMEMS-NWS-QUID-  
1133 004-002 report.
- 1134 Nowack, P., Braesicke, P., Haigh, J., Abraham, N. L., Pyle, J., and Voulgarakis, A. (2018). Us-  
1135 ing machine learning to build temperature-based ozone parameterizations for climate sensitivity  
1136 simulations. *Environmental Research Letters*, 13(10):104016.
- 1137 Ouala, S., Fablet, R., Herzet, C., Chapron, B., Pascual, A., Collard, F., and Gaultier, L. (2018).  
1138 Neural network based Kalman filters for the spatio-temporal interpolation of satellite-derived sea  
1139 surface temperature. *Remote Sensing*, 10(12):1864.
- 1140 Pingree, R. and Griffiths, D. (1978). Tidal fronts on the shelf seas around the British Isles. *Journal  
1141 of Geophysical Research: Oceans*, 83(C9):4615–4622.
- 1142 Pradhan, H. K., Völker, C., Losa, S. N., Bracher, A., and Nerger, L. (2020). Global assimilation  
1143 of ocean-color data of phytoplankton functional types: Impact of different data sets. *Journal of  
1144 Geophysical Research: Oceans*, 125(2):e2019JC015586.
- 1145 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f.  
1146 (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*,  
1147 566(7743):195–204.

- 1148 Sacco, M. A., Pulido, M., Ruiz, J. J., and Tandeo, P. (2024). On-line machine-learning forecast un-  
 1149 certainty estimation for sequential data assimilation. *Quarterly Journal of the Royal Meteorological*  
 1150 *Society*, 150(762):2937–2954.
- 1151 Sacco, M. A., Ruiz, J. J., Pulido, M., and Tandeo, P. (2022). Evaluation of machine learning tech-  
 1152 niques for forecast uncertainty quantification. *Quarterly Journal of the Royal Meteorological Society*,  
 1153 148(749):3470–3490.
- 1154 Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., Ward, B. A., Slawig,  
 1155 T., and Oschlies, A. (2017). Reviews and syntheses: parameter identification in marine planktonic  
 1156 ecosystem modelling. *Biogeosciences*, 14(6):1647–1701.
- 1157 Schmidtke, S., Stramma, L., and Visbeck, M. (2017). Decline in global oceanic oxygen content during  
 1158 the past five decades. *Nature*, 542(7641):335–339.
- 1159 Shulman, I., Frolov, S., Anderson, S., Penta, B., Gould, R., Sakalaukus, P., and Ladner, S. (2013). Im-  
 1160 pact of bio-optical data assimilation on short-term coupled physical, bio-optical model predictions.  
 1161 *Journal of Geophysical Research: Oceans*, 118(4):2215–2230.
- 1162 Simon, E. and Bertino, L. (2012). Gaussian anamorphosis extension of the DEnKF for combined state  
 1163 parameter estimation: Application to a 1D ocean ecosystem model. *Journal of Marine Systems*,  
 1164 89(1):1–18.
- 1165 Simon, E., Samuelson, A., Bertino, L., and Mouysset, S. (2015). Experiences in multiyear combined  
 1166 state–parameter estimation with an ecosystem model of the North Atlantic and Arctic Oceans using  
 1167 the ensemble Kalman filter. *Journal of Marine Systems*, 152:1–17.
- 1168 Skakala, J., Awty-Carroll, K., Menon, P. P., Wang, K., and Lessin, G. (2023). Future digital twins:  
 1169 emulating a highly complex marine biogeochemical model with machine learning to predict hypoxia.  
 1170 *Frontiers in Marine Science*, 10:1058837.
- 1171 Skakala, J., Bruggeman, J., Brewin, R. J., Ford, D. A., and Ciavatta, S. (2020). Improved represen-  
 1172 tation of underwater light field and its impact on ecosystem dynamics: A study in the North Sea.  
 1173 *Journal of Geophysical Research: Oceans*, 125(7):e2020JC016122.
- 1174 Skákala, J., Ford, D., Brewin, R. J., McEwan, R., Kay, S., Taylor, B., de Mora, L., and Ciavatta,  
 1175 S. (2018). The assimilation of phytoplankton functional types for operational forecasting in the  
 1176 Northwest European shelf. *Journal of Geophysical Research: Oceans*, 123(8):5230–5247.
- 1177 Skákala, J., Ford, D., Bruggeman, J., Hull, T., Kaiser, J., King, R. R., Loveday, B., Palmer, M. R.,  
 1178 Smyth, T., Williams, C. A., et al. (2021). Towards a multi-platform assimilative system for North  
 1179 Sea biogeochemistry. *Journal of Geophysical Research: Oceans*, 126(4):e2020JC016649.
- 1180 Skákala, J., Ford, D., Fowler, A., Lea, D., Martin, M. J., and Ciavatta, S. (2024). How uncertain and  
 1181 observable are marine ecosystem indicators in shelf seas? *Progress in Oceanography*, 224:103249.
- 1182 Smith, V. H. and Schindler, D. W. (2009). Eutrophication science: where do we go from here? *Trends*  
 1183 *in ecology & evolution*, 24(4):201–207.
- 1184 Smyth, T. J., Fishwick, J. R., Al-Moosawi, L., Cummings, D. G., Harris, C., Kitidis, V., Rees, A.,  
 1185 Martinez-Vicente, V., and Woodward, E. M. (2010). A broad spatio-temporal view of the Western  
 1186 English Channel observatory. *Journal of Plankton Research*, 32(5):585–601.
- 1187 Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J. (2016). Data assimilation in a coupled  
 1188 physical–biogeochemical model of the California current system using an incremental lognormal 4-  
 1189 dimensional variational approach: Part 1—model formulation and biological data assimilation twin  
 1190 experiments. *Ocean Modelling*, 106:131–145.
- 1191 Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V. (2021). Bridging  
 1192 observations, theory and numerical simulation of the ocean using machine learning. *Environmental*  
 1193 *Research Letters*, 16(7):073008.
- 1194 Sonntag, S. and Hense, I. (2011). Phytoplankton behavior affects ocean mixed layer dynamics through  
 1195 biological-physical feedback mechanisms. *Geophysical Research Letters*, 38(15).

- 1196 Telszewski, M., Palacz, A., and Fischer, A. (2018). Biogeochemical in situ observations—motivation,  
1197 status, and new frontiers. *New Frontiers in Operational Oceanography*, pages 131–160.
- 1198 Teruzzi, A., Bolzon, G., Feudale, L., and Cossarini, G. (2021). Deep chlorophyll maximum and  
1199 nutricline in the Mediterranean Sea: emerging properties from a multi-platform assimilated biogeo-  
1200 chemical model experiment. *Biogeosciences*, 18(23):6147–6166.
- 1201 Teruzzi, A., Dobricic, S., Solidoro, C., and Cossarini, G. (2014). A 3-D variational assimilation scheme  
1202 in coupled transport-biogeochemical models: Forecast of Mediterranean biogeochemical properties.  
1203 *Journal of Geophysical Research: Oceans*, 119(1):200–217.
- 1204 Umlauf, L. and Burchard, H. (2011). Diapycnal transport and mixing efficiency in stratified boundary  
1205 layers near sloping topography. *Journal of physical oceanography*, 41(2):329–345.
- 1206 Vagle, S., McNeil, C., and Steiner, N. (2010). Upper ocean bubble measurements from the NE Pacific  
1207 and estimates of their role in air-sea gas transfer of the weakly soluble gases nitrogen and oxygen.  
1208 *Journal of geophysical research: oceans*, 115(C12).
- 1209 van der Merwe, R., Leen, T. K., Lu, Z., Frolov, S., and Baptista, A. M. (2007). Fast neural network  
1210 surrogates for very high dimensional physics-based models in computational oceanography. *Neural  
1211 Networks*, 20(4):462–478.
- 1212 Wakelin, S. L., Artioli, Y., Butenschön, M., Allen, J. I., and Holt, J. T. (2015). Modelling the  
1213 combined impacts of climate change and direct anthropogenic drivers on the ecosystem of the  
1214 Northwest European continental shelf. *Journal of Marine Systems*, 152:51–63.
- 1215 Wakelin, S. L., Artioli, Y., Holt, J. T., Butenschön, M., and Blackford, J. (2020). Controls on near-  
1216 bed oxygen concentration on the Northwest European Continental Shelf under a potential future  
1217 climate scenario. *Progress in Oceanography*, 187:102400.