

Hybrid machine learning data assimilation for marine biogeochemistry

Ieuan Higgs^{1,2}, Ross Bannister^{1,2}, Jozef Skákala^{2,3}, Alberto Carrassi^{1,4}, and Stefano Ciavatta⁵

¹Department of Meteorology, University of Reading, UK

²National Centre for Earth Observation, UK

³Plymouth Marine Laboratory, UK

⁴Department of Physics and Astronomy “Augusto Righi”, University of Bologna, IT

⁵Mercator 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, ~~strong nonlinearity~~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 ~~significantly enhances~~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 ~~moderate~~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, ~~optimizing~~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
54 only a small fraction of the model variables. In the context of operational marine BGC forecasting,
55 these observations are typically satellite ocean colour-derived chlorophyll (Fennel et al., 2019; Groom
56 et al., 2019), ~~with~~. Furthermore the assimilation of BGC-Argo observations (including chlorophyll,
57 nitrate and oxygen) in open ocean waters has been recently implemented in state-of-the-art operational
58 systems (Cossarini et al., 2019; Teruzzi et al., 2021). Other products are assimilated in reanalyses
59 or research and development (R&D) versions of the operational systems, such as optical variables
60 (Shulman et al., 2013; Ciavatta et al., 2014; Jones et al., 2016; Gregg and Rousseaux, 2017; Skakala
61 et al., 2020) and size-class chlorophyll (Ciavatta et al., 2018, 2019; Skákala et al., 2018; Pradhan et al.,
62 2020), as well as types of in situ data, such as chlorophyll, oxygen and nutrients from gliders (Skákala
63 et al., 2021). For a broader range of marine BGC DA work beyond operational applications, see many
64 other references, e.g., Simon and Bertino (2012); Shulman et al. (2013); Gehlen et al. (2015); Simon
65 et al. (2015).

66 Different DA systems are used across marine BGC forecasting centres, including variational (Ford
67 et al., 2012; Song et al., 2016; Skákala et al., 2018; Coppini et al., 2021), Singular Evolutive Extended
68 Kalman filter (SEEK) (Gutknecht et al., 2019; Ciliberti et al., 2021) and Ensemble Kalman Filter
69 (EnKF) (Bertino et al., 2021) -based methods. Although ensemble methods (e.g., EnKF) are appeal-
70 ing for their capability to provide uncertainty quantification and cross-covariances, the more complex
71 marine BGC models such as the European Regional Seas Ecosystem Model (ERSEM) (Butenschön
72 et al., 2016) or the Biogeochemical Flux Model (BFM) (Cossarini et al., 2017), currently rely on
73 variational methods as running a sufficiently large ensemble in the day-to-day operational forecasting
74 context can be computationally expensive. Moreover, for such complex models, variational methods
75 update only a very limited number of unobserved variables, typically using very simple balancing prin-
76 ciples based on the simulated structure and stoichiometry of the phytoplankton community (Teruzzi
77 et al., 2014; Skákala et al., 2018). We will call such systems with certain approximations “univariate”,
78 and systems that update (nearly) all model state variables as a direct result of DA “multivariate”.
79 The multivariate updates can happen in the DA step, through ensemble-informed background¹ error
80 covariances (as in the EnKF), through balancing schemes, such as the scheme of Hemmings et al.
81 (2008) based on nitrogen mass conservation applied to Nutrient-Phytoplankton-Zooplankton-Detritus
82 models (Hemmings et al., 2008; Ford et al., 2012), or through the tangent-linear and adjoint models
83 Mattern et al. (2017). However, whenever such multivariate schemes were applied to highly complex
84 marine BGC models (in reanalyses, or R&D), the improvement ~~on~~-of non-observed variables was
85 typically marginal, with several variables often systematically degraded by DA (e.g., Ciavatta et al.
86 (2016, 2018)). This provides a warning on the use of incorrect assumptions in multivariate balancing
87 schemes or in the EnKFs and the need for better DA and/or ensemble design.

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

99 In this work, we investigate ~~the capability of ML to learn the whether ML can capture the complex,~~
100 ~~non-linear /relationships among biogeochemical (BGC) variables, and how these relationships depend~~
101 ~~on the evolving state of the system (i.e., are flow-dependentrelations between BGC variables, before~~
102 ~~using those learned functions-). Here, flow-dependent indicates that the statistical relationships or~~
103 ~~errors between variables vary with the current flow or dynamical state, rather than being constant or~~
104 ~~solely climatological. The relationships learned by the ML model are then used~~ within a DA scheme ~~or~~
105 ~~to fully substitute, or as a full substitute for~~ it. Thus, we are not attempting to directly emulate or

¹In 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 improve BGC models via ML, but instead use ML to improve DA, and specifically to cope with the
107 challenging problem of propagating information from observed to unobserved variables in single-model
108 deterministic runs that, by definition, do not have an ensemble to estimate the necessary statistics.
109 The main goal of this approach is to introduce multivariate DA into the system, whilst benefiting
110 from the relatively low computational cost of ML. This study falls within a stream of research aimed
111 at building suitable hybrid ML-DA schemes (see Buizza et al., 2022; Cheng et al., 2023, and references
112 therein), and, to our knowledge, it is the first such attempt in the context of marine BGC.

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

119 In a second configuration, instead of merging DA and ML, ~~the former DA~~ is used to ~~produce~~
120 generate a training dataset, from which ML learns the full DA step ~~, in an “in an” end-to-end” fashion~~
121 ~~(Barth et al., 2020; Fablet et al., 2021).~~ ~~Here, the ML task is that of DA as a whole, i.e., given the~~
122 ~~background state and observations, return the analysis increments to”~~ emulation (Barth et al., 2020; Fablet et al., 2021)
123 . In this context, “end-to-end” refers to a ML model trained to map inputs (such as forecast states
124 and observational information) directly to analysis increments, bypassing traditional, hand-crafted
125 intermediate steps. As such, the background ML model learns to predict the analysis increments for
126 unobserved variables. ~~As mentioned above, we do not intend substituting/improving the BGM model,~~
127 ~~and our,~~ given the surface background state and the increment for the observed variable. Recent
128 work by Bocquet et al. (2024) has demonstrated that the entire EnKF analysis step can be learned
129 in this data-driven, end-to-end learning focuses only on learning the instantaneous DA updates while
130 using the given BGC model to issue the forecasts. Efficient end-to-end learning of the EnKF analysis
131 in chaotic systems has been recently proven by Bocquet et al. (2024) manner.

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

137 Our work has a potentially important application within the North-West European Shelf (NWES)
138 operational DA system to which it is tailored. Yet we will discuss its generalisation to other compa-
139 rable systems, applied to spatial domains with similar type of marine BGC dynamics. Based on the
140 transferability of the ML model, we speculate whether it is feasible to use the ML model trained in
141 1D on a 3D domain and propose a methodology for doing so.

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

150 2 Model and data assimilation setups for biogeochemistry

151 2.1 Physical model: GOTM

152 The Generalised Ocean Turbulence Model (GOTM) (Bolding and Villarreal, 1999) is a 1D water
153 column model for studying hydrodynamic and biogeochemical processes when coupled to a biogeo-
154 chemical model, in marine and limnic waters. ~~It provides~~ GOTM provides the necessary physical
155 forcing for the coupled biogeochemical model, offering a balance between realism and computational
156 cost by using real atmospheric forcing data, relaxation profiles, and coupling at full BGC complexity,
157 while sacrificing the explicit representation of 3D processes. This makes the system ideal for this
158 work, where we are primarily interested in the error relationships between different biogeochemical
159 quantities (e.g., chlorophyll to nitrate), rather than the spatial error characteristics. GOTM can be
160 used as a stand-alone model for studying dynamics of boundary layers in natural waters, having hy-
161 drodynamic applications in investigations of air-sea fluxes (Vagle et al., 2010), surface mixed-layer
162 dynamics (Sonntag and Hense, 2011), dynamics of bottom boundary layers with or without sediment

163 transport (Umlauf and Burchard, 2011; Falchetti et al., 2010), and estuarine and coastal dynamics
 164 (Burchard, 2009).

165 2.2 Biogeochemical model: ERSEM

166 ERSEM (Baretta et al., 1995; Butenschön et al., 2016) is a marine biogeochemistry model that sim-
 167 ulates lower trophic levels of the ocean ecosystem, including plankton and benthic fauna (Blackford,
 168 1997), see Table 1. The model divides phytoplankton into four functional types based on size: pico-
 169 phytoplankton, nanophytoplankton, microphytoplankton and diatoms (Baretta et al., 1995). ERSEM
 170 uses variable stoichiometry for the simulated plankton groups (Baretta-Bekker et al., 1997; Geider
 171 et al., 1997) and represents the biomass of each functional type in terms of chlorophyll, carbon, ni-
 172 trogen, and phosphorus, with diatoms also being represented by silicon. ERSEM predators consist of
 173 three types of zooplankton (mesozooplankton, microzooplankton, and heterotrophic nanoflagellates),
 174 with organic material being decomposed by a single type of heterotrophic bacteria (Butenschön et al.,
 175 2016). The model represents three different sizes of detritus (small, medium and large) and three
 176 types of dissolved organic matter (DOM: refractory; semi-labile; labile). The inorganic component
 177 of ERSEM includes nutrients such as nitrate, phosphate, silicate, ammonium, and carbon, as well as
 178 dissolved oxygen. The carbonate system is also included in the model (Artioli et al., 2012). ERSEM
 179 has been used for many applications including NWES and Mediterranean Sea biogeochemistry reanal-
 180 yses (Ciavatta et al., 2016, 2018, 2019), NWES operational ~~forecast~~ [forecasts](#) (Skákala et al., 2018;
 181 McEwan et al., 2021), and NWES climate projections (Wakelin et al., 2015, 2020; Galli et al., 2024).

Functional Group	Class/Type	Chemical Components
Phytoplankton	Diatoms	$\text{Chl}_{\chi_{\text{dia}}}$, C, N, P, Si
Functional Types (PFT)	Microphytoplankton	$\text{Chl}_{\chi_{\text{micro}}}$, C, N, P
	Nanophytoplankton	$\text{Chl}_{\chi_{\text{nano}}}$, C, N, P
	Picophytoplankton	$\text{Chl}_{\chi_{\text{pico}}}$, 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: Chl_{χ} is chlorophyll; C is carbon; N is nitrogen; P is phosphorus and Si is silicon. Note that we also use total chlorophyll (denoted as ~~e-chl~~ in this [paperwork](#)), which is a diagnostic variable calculated as the sum of chlorophyll concentrations from all PFT classes.

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

2.3 Model configuration and synthetic data setup

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

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

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

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

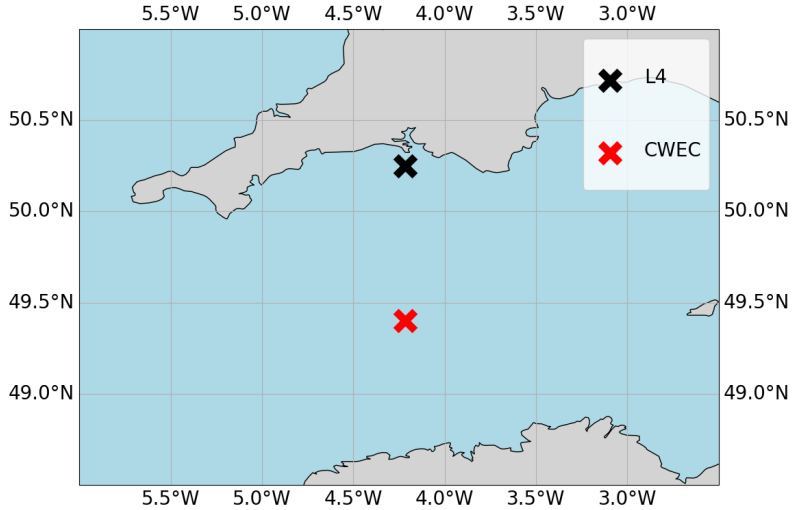


Figure 1: Map of the Western English Channel, marking the L4 model-training location with a black cross and the CWEC (Central Western English Channel) with a red cross, where we evaluated the model portability.

227 For the purposes of training ML models, and generating climatological statistics, time periods for
 228 the L4 location are partitioned as follows: training data (2000-2014), validation (2015-2017), offline
 229 test (2018-2020), online test (2022-2023). Offline refers here to a setup in which the ML-OI-DA
 230 analysis is not then used-cycled as the initial condition for the next forecast, and so it does not
 231 impact successive DA cycles. Conversely, online-Online refers to a setup in which updates-to-the
 232 system-assimilation updates are cycled, and so can have dynamical impact on later DA cycles as the
 233 model integrates forward in time. Climatological correlations and variances are calculated using a
 234 free-run ensemble over the training period. Because the forecasts extends seven days into the future,
 235 the number of available samples for any specific calendar day was limited, even though forecasts were
 236 issued throughout the full training period. To address this, the climatological statistics were computed
 237 as daily values, defined by averaging the statistics for a given day across all years in the training set.
 238 To increase the sample size and smooth out variability, a ± 30 -day window around each calendar day
 239 was applied. The CWEC location uses a run spanning 2000 - 2010 to generate climatological statistics,
 240 and the online test is performed for (2022-2023). No validation or offline test period is required for
 241 CWEC in this work, as we are primarily interested in a “naive transferability” of the model trained
 242 and evaluated at L4.

2.4 Data assimilation setups

243 We examine a total of five data assimilation (DA) setups in this work. ~~These are conventional DA~~
 244 ~~methods—namely:~~ a simple univariate scheme to reflect how DA is done currently in representative
 245 of current operational marine BGC systems, and a scheme that uses climatological statistics to update
 246 unobserved variables, an EnKF for comparison ~~to the new schemes that~~, and two novel hybrid DA
 247 schemes which are hybridised with ML techniques. Before describing each scheme, we give the basic
 248 equations of conventional DA, introduce the state vector-vectors that the DA uses, the observation
 249 type that we will assimilate, and other adjustments that are done post assimilation.

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

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

252 and where \mathbf{K} is the Kalman gain matrix

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

253 and where \mathbf{x}^a is the analysis (updated) state, \mathbf{x}^b is the background state (which in this work is
 254 equivalent to the state of the system after a seven-day forecast stateperiod), \mathbf{y} are the observations, \mathcal{H}

256 is the observation operator (with Jacobian \mathbf{H}), \mathbf{P}^b is the background error covariance matrix, and \mathbf{R}
 257 is the observation error covariance matrix. The matrix \mathbf{P}^b is of special interest to this work. Ideally
 258 this matrix should be appropriately flow-dependent, but in practice it is often not, such as in many
 259 operational schemes ~~where a fixed climatological estimate is used. For instance, in marine BGC, we~~
 260 ~~would expect the onset of a phytoplankton bloom to trigger changes in nutrient concentrations, which~~
 261 ~~in turn affect the correlations between the two. This response cannot be captured by climatological~~
 262 ~~statistics alone.~~ The purpose of this work is to introduce such flow-dependency to \mathbf{P}^b , or to the
 263 analysis increments $\Delta\mathbf{x}$, with ML techniques. ~~While each of the DA schemes described later will~~
 264 ~~use different methods to update unobserved variables, all schemes will update total chlorophyll, the~~
 265 ~~observed variable, and it's associated PFTs in the same way (using the RUS scheme described in~~
 266 ~~Sect. 2.4.1).~~

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

285 We assimilate only ~~observations of~~ total chlorophyll, y , at the surface. Total chlorophyll in the
 286 model, ~~x_{chl}~~ x_{chl} , is a diagnostic variable obtained by summing the chlorophyll content of all phyto-
 287 plankton functional types (PFTs, see Table 1), which are themselves prognostic variables. In principle,
 288 one could keep only the PFT chlorophyll concentrations in the state and represent total chlorophyll
 289 via the observation operator. However, this would require explicitly summing across PFTs at every
 290 assimilation step, making the DA equations more cumbersome. Instead, we treat surface total chloro-
 291 phyll directly as part of the state vector. This simplifies the presentation of the analysis equations,
 292 since the observation operator then reduces to a simple selection operator:

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

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

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

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

300 where ~~x_{χ_i}~~ x_{χ_i} stands for the chlorophyll component of each PFT, ~~i~~ , ~~as in Table 1~~.

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

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

303 where ~~$\zeta_{i,j}$~~ $\zeta_{i,j}$ represents the non-chlorophyll components, ~~j~~ , of each PFT, ~~i~~ . This constrained re-
 304 distribution scheme (Teruzzi et al., 2014; Skákala et al., 2018) ensures that phytoplankton updates

305 preserve forecast stoichiometric ratios and remain physiologically consistent, rather than allowing **the**
306 **EnKF—a statistical method (such as the EnKF)** to update PFTs freely. While **we would expect** an
307 unconstrained EnKF ~~would to~~ eventually converge towards similar balances, this explicit approach
308 guarantees that assimilation respects acclimation dynamics and maintains the community structure
309 of the model. Importantly, the same ratio-based balancing scheme can also be applied outside an
310 ensemble framework in single-model runs, where it provides a consistent way of updating PFTs from
311 a total chlorophyll correction.

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

Run / scheme	Description / purpose	<i>chl</i> variance	<i>i</i> variance	<i>chl-i</i> cor- relation source	Δx_{chl}	Δx_i
Preparation / training runs						
1. Truth run	To synthesise observations and for analysis evaluation	n/a	n/a	n/a	n/a	n/a
2. Ensem- ble 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)
Conventional assimilation runs						
4. RUS	Reference univariate scheme (TC + stoichiometrical PFT update) / baseline for extensions	climatology	n/a	n/a	Eq. (6)	zero
RUS extension assimilation runs (update to variable <i>i</i> with ML methods)						
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
RUS extension assimilation runs (update to variable <i>i</i> with non-ML methods)						
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).

2.4.1 Reference univariate DA scheme (RUS)

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

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

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

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

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

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 (~~potentially~~ $1 \leq i \leq N$) have updates according to

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

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 σ_i and σ_{chl} are their respective background error standard deviations. In Eq. (7) the analysis increment of the update is labelled.

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

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

369 Each ML approach is tested in the following scenarios: (1) a set-up where the only unobserved
370 variable that is updated is nitrate, (2a) a set-up where we update the full set of pelagic variables,
371 and (2b) a set-up where we update a partial set of the pelagic variables, eliminating poorly estimated
372 variables based on the results of (2a). The progression from (1) to (2a) allows us to move from a
373 controlled test of a single key limiting variable to a comprehensive update of the full system, while
374 (2b) represents a refinement step that is only possible after evaluating the performance of (2a). In this
375 way, the experimental design not only tests the limits of updating all variables, but also demonstrates
376 how excluding problematic variables can improve robustness without discarding the broader benefits
377 of multivariate updates. In each setup, the number of outputs to be estimated by the ML model
378 corresponds to the number of unobserved variables. In the case of ML-OI (see Sect. 3.1), the standard
379 deviations must also be estimated from climatology.

380 3.1 Hybrid machine-learning optimal interpolation (ML-OI)

381 This approach first updates the observed total chlorophyll and associated PFTs in an identical manner
382 to the RUS described in Sect. 2.4.1. Then, an ANN estimates the state-dependent correlations $\rho_{i,chl}$
383 between observed and unobserved quantities in Eq. (8) as a function of the background state (even
384 though the EnKF update of state variables from other state variables is linear, the relationship between
385 state variables and correlations is likely non-linear). Together with climatologically estimated values
386 of σ_i and σ_{chl} (estimated using a free-run ensemble over the training period as described in Sec. 2.3),
387 the correlations are substituted into Eqs. (8) and then (7) to provide updates to the unobserved vari-
388 ables in the system. We call this approach ML-OI (“optimal interpolation”, Table 2, row 5). For each
389 variable input into the ML-OI model to estimate the correlation, the background state is additionally
390 divided by its climatological maximum at the corresponding location (before the regular standardisa-
391 tion procedure is applied to the data). This ~~normalization~~normalisation accounts for differences in the
392 amplitude of seasonal variability while making ~~a bold an~~ assumption that the underlying correlative
393 relationships between variables remain consistent across locations. Since correlations are dimension-
394 less, this scaling does not affect their interpretation, and the subsequent analysis increments (which
395 are constructed by combining the estimated correlations with the location-specific climatological vari-
396 ances) remain physically consistent. ~~Since variances are~~Apart from correlation, the other aspect of
397 covariance to determine are the variances. Since each variance is a single-variable ~~statistics~~statistic
398 that can be estimated more reliably than correlations from long climatological records, we assume they
399 are sufficiently robust to provide a stable basis for use in the Kalman gain. In contrast, correlations
400 describe joint variability and therefore require the more sophisticated, data-driven approach described
401 above, and cannot be approximated in the same way.

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

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

411 While the field of hybrid ML-DA is growing rapidly, there exists relatively few works in which ML-
412 estimated background error covariances are so closely coupled with existing DA systems. However,
413 a few particularly relevant examples stand out such as Ouala et al. (2018), in which a Kalman-like
414 analysis update is applied to satellite-derived sea surface temperature fields using ~~artificial neural~~

415 ~~network (ANN)-estimated~~ ANN-estimated background error covariances. Additional examples of this
416 can be seen in Sacco et al. (2022), which aim to learn different sources of uncertainty using ANNs
417 on both toy models and sea level pressure forecasts. Further work (Sacco et al., 2024) uses an EnKF
418 to generate flow dependent background error covariances, and then learns them using a convolutional
419 neural network.

420 3.2 End-to-end machine learning of EnKF updates (ML-EtE)

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

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

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

443 This approach follows other non-marine BGC work in a similar direction, such as Bonavita and
444 Laloyaux (2020), who used ANNs to emulate the main features of an operational weak-constraint
445 4D-Var scheme, while Bocquet et al. (2024) pursued a similar end-to-end replacement of the analysis
446 step. Likewise, several studies have demonstrated the emulation of analysis increments to estimate
447 and correct model error (Brajard et al., 2020; Gregory et al., 2024). A common feature of these
448 works is their reliance on an existing DA system or, more generally, on a robust reanalysis. The
449 need for such a reanalysis represents a key limitation of this approach ~~—~~ one we revisit in the
450 conclusion. Here, however, our primary goal is to examine the feasibility of ML-EtE and its ability to
451 learn the EnKF updates effectively. Using this approach, the increments still ultimately come from
452 the “analysis–background” of an EnKF, which provides a linear update to the system (even if the
453 relationship between the state at the analysis increments is non-linear). Going beyond this limitation
454 would require either training on increments relative to the true state (e.g., truth–background) or
455 employing a non-linear DA system, such as a particle filter, to capture more complex, non-linear
456 corrections.

457 3.3 Purely climatological updates

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

462 3.4 Skill metric and machine learning model evaluation

463 3.4.1 Skill metric

464 For a system that runs for τ cycles (where a cycle represents a complete 7-day forecast and analysis),
465 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 .
466 The expected RMSE (root mean square error) over M ensemble members (or a set of M single-model

467 runs), is calculated as:

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

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

471 3.4.2 SHAP analysis

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

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

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

498 4 Results and discussion

499 4.1 System dynamics

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

- 507 • The *light-limited regime* ~~typically describes a fully mixed water column, which~~ approximately
508 spans the period from October to the start of the next spring bloom. Here, there is little-to-no
509 phytoplankton growth due to the reduced light-levels at this time of year, meaning nutrients ~~are~~
510 can be mixed throughout the water column without being used by the phytoplankton. During
511 this period, phytoplankton concentrations are very low and mostly decoupled from nutrient
512 dynamics ².

²The 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.

- 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.

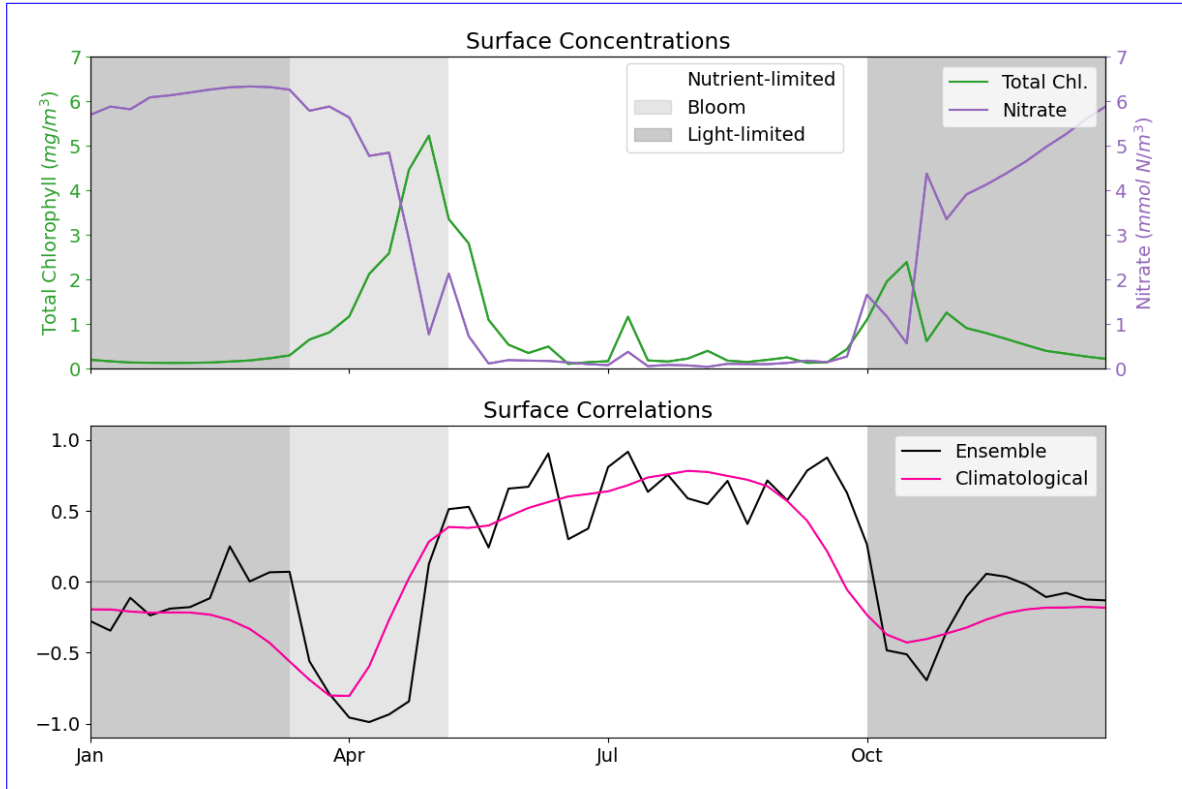


Figure 2: The top panel shows a time series for the surface concentrations of total chlorophyll (green) and nitrate (purple) during 2023 at the L4 location. The bottom panel presents correlations between total chlorophyll and nitrate derived from a 96-member ensemble (black) and from a daily varying 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 ~~us a more understandable the simplest~~ proof-of-concept ~~with reduced complexity 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 ~~See Sect.~~ 4.3, this strategy for assigning importance ~~or priority~~ to variables in the DA scheme does not necessarily ~~correspond to the dynamical importance of a variable~~ ~~reflect their true dynamical role~~ in the model, highlighting the need ~~for specific results to evaluate how different assimilation strategies and variable-update choices affect performance.~~

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

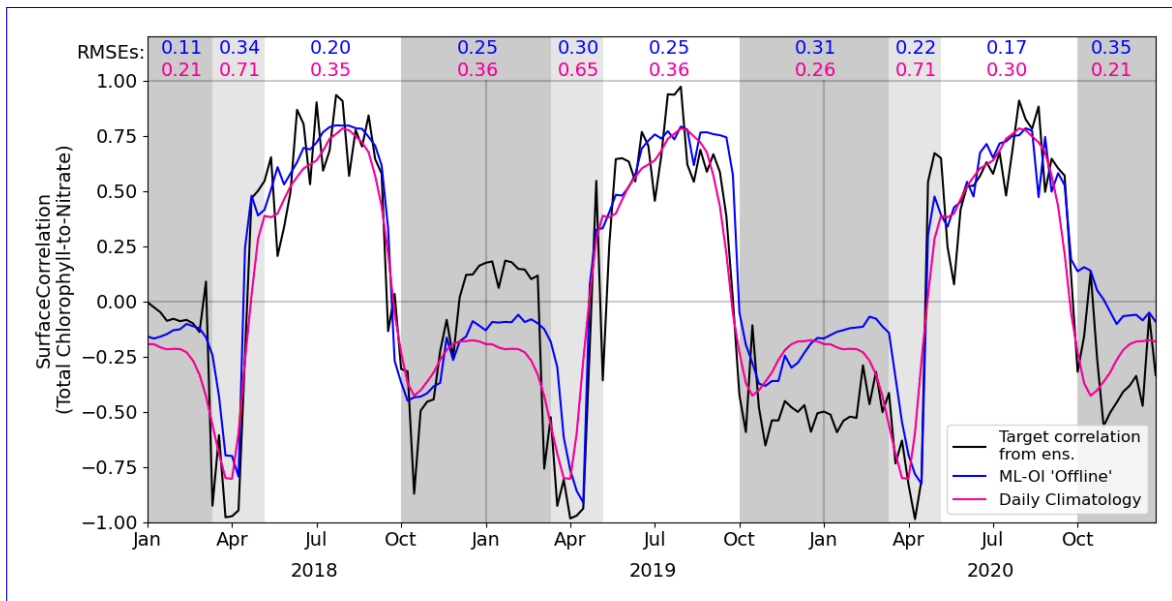


Figure 3: Estimates of correlation between total chlorophyll and nitrate, at weekly intervals across the 3-year offline test period. The target correlation (black) is calculated from the 100-member free-run ensemble (Table 2, row 2). Correlation estimates are shown for ML-OI (blue) and the daily climatology 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.

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

548 Most clearly, ML-OI is better than climatology for estimating the highly distinctive correlative pat-
 549 tern between total chlorophyll and nitrate during the bloom regime, showing a moderate to significant
 550 RMSE reduction in every bloom period. This pattern consists of a sharp drop to a strongly negative
 551 correlation, before an almost instantaneous increase to a strong positive correlation. These correlation
 552 patterns can be simply explained. During the bloom, phytoplankton growth exhausts nutrients, lead-
 553 ing to negative correlations between chlorophyll and nutrients, whilst the end of the bloom, and the
 554 following period, phytoplankton growth is nutrient limited, leading to positive correlation. The precise
 555 timing of the bloom (and hence this correlation pattern) has notable inter-annual variability – vary-
 556 ing within a period of approximately 5 weeks each year, in this model. The climatological correlations
 557 estimate this pattern poorly as they are smoothed over this period of inter-annual variability, but the
 558 ML-OI model, which estimates correlations from the state of the marine BGC model, captures the
 559 pattern much more accurately.

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

567 Finally, we also see that during the light-limited regime, the system can exist in either a “weakly
 568 positive or no correlation” state, or a “moderately negative correlation” state. However, both the
 569 climatological and ML ~~estimate estimates~~ fail to capture these possible states (both predicting a weakly

570 negative correlation over the period). During this time, total chlorophyll and nitrate are generally
 571 decoupled, and there is ~~there is~~ no clear link between the state of the system and the correlations
 572 estimated. Furthermore, as the ~~ensemble concentrations in concentrations of~~ total chlorophyll are very
 573 near zero ~~, the at this time of year (both in the ensemble and observations) and there is little spread~~
 574 in the ensemble – the resulting DA updates are ~~also~~ small and have very little impact. This means
 575 that any improvement or degradation in correlation estimates at this time of year are less likely to
 576 result in any great improvement to the system, as there is weak relationship between chlorophyll and
 577 nitrate DA increments and the updates are small. However, updates at the start of this period can be
 578 important, as the resulting store of nitrate in the upper water column could have dynamical impact
 579 in later DA cycles when light is no longer limiting and the next bloom period starts.

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

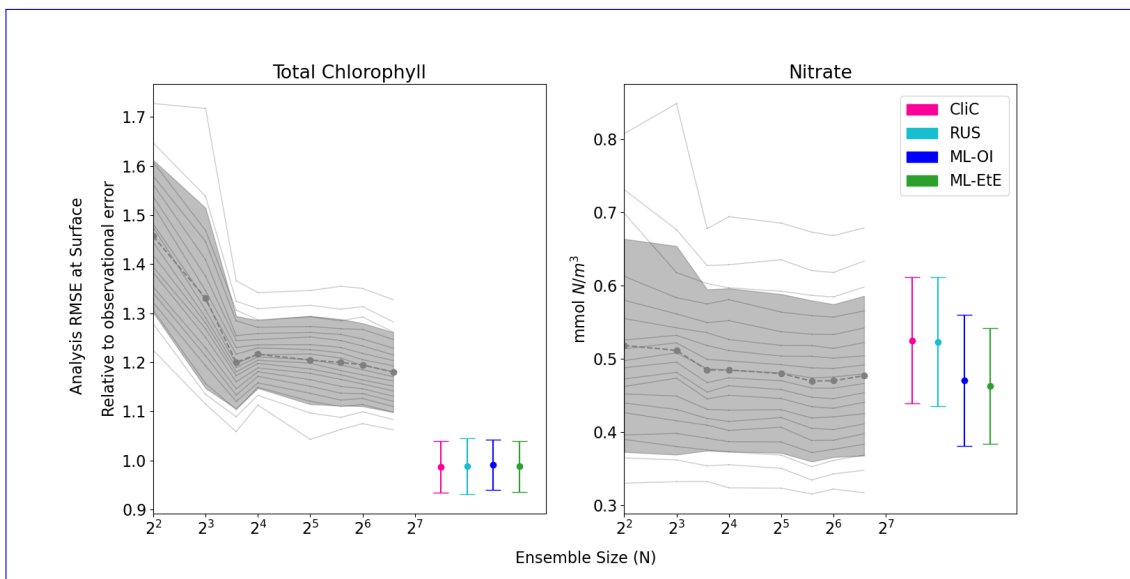


Figure 4: The relationship between analysis RMSE (Eq. (9)) and ensemble size for EnKFs with different ensemble sizes, as well as the performance of the different single-model run schemes. The left panel shows the RMSE of the observed variable, total chlorophyll, normalised relative to the observational 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 ± 1 standard deviation of ensemble member error. The mean error and ± 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.

585 Figure 4 displays the performance of the EnKF, the RUS scheme, the climatological statistics
 586 scheme (Clic), and the ML schemes. Each panel shows the mean expected error of ensemble members
 587 for ensemble sizes ranging from 4 to 96. In the left panel for total chlorophyll, the relative RMSE is
 588 calculated as a ratio of the observation error. The EnKF achieves ~~a~~ near-optimal performance ~~at an~~
 589 ensemble size > 16 , after for ensemble sizes greater than 16, after which the mean expected error ~~of~~
 590 ensemble members reaches a plateau with increasing size. The relative analysis error of total chlorophyll
 591 ~~is normalised according to observational error. This exceeds a value of 1 as we are measuring because~~
 592 it is based on the expected error of ensemble members, not ~~error to~~ the ensemble mean ~~, as described~~
 593 in (see Sect. 3.4.1). This reflects the additional stochastic error introduced by the EnKF’s generation
 594 of perturbed observations. Since the EnKF generates an ensemble of observations (with noise based
 595 on the uncertainty), an additional source of error is introduced relative to the single model runs which
 596 are not stochastic. This means that when we calculate the error of each ensemble member, rather
 597 than the error of the ensemble mean, we get this difference in error. We also see, in the right panel,
 598 that the error decreases with ensemble size for the unobserved nitrate, indicating that the system

599 converges towards more correct nitrate updates at larger ensemble sizes.

600 As expected, the analysis error in the observed total chlorophyll is generally comparable across
601 each scheme because they all use the same method, the RUS scheme of Sect. 2.4.1, to update the
602 observed total chlorophyll. However, there are more noticeable differences in the schemes that extend
603 the updates to nitrate as well. ~~In this, we~~ We can clearly see that both the RUS scheme (no update to
604 nitrate) and the CliC scheme (update of nitrate using climatological covariances in Eq. (8)) perform
605 similarly poorly ~~in improving analysis error of for~~ nitrate – meaning that the information provided
606 by the observation has not propagated well to the unobserved variable. In contrast to this, both ML
607 approaches result in a significant improvement in performance, reducing analysis error by between 8 –
608 12%. This means that the information from observations can effectively propagate to the unobserved
609 variables in a single-model run, without the need for an expensive ensemble to model the statistics at
610 run time. ~~Also, this indicates that the improvement in correlation estimation shown in ML-EtE shows~~
611 ~~similar improvements during the online testing period as those observed in~~ the offline experiment ~~of~~
612 ~~(Fig. 3 translates (at least on average) into the online testing period, where the updates of a given)~~;
613 ~~indicating that the benefits persist even when the updates from each~~ DA cycle feed into subsequent
614 ~~cycles. The standard deviation of ones. Moreover, ML-EtE is also lower than the spread for exhibits~~
615 ~~a lower standard deviation than~~ ML-OI, ~~which is a good sign of lower indicating reduced~~ sensitivity.

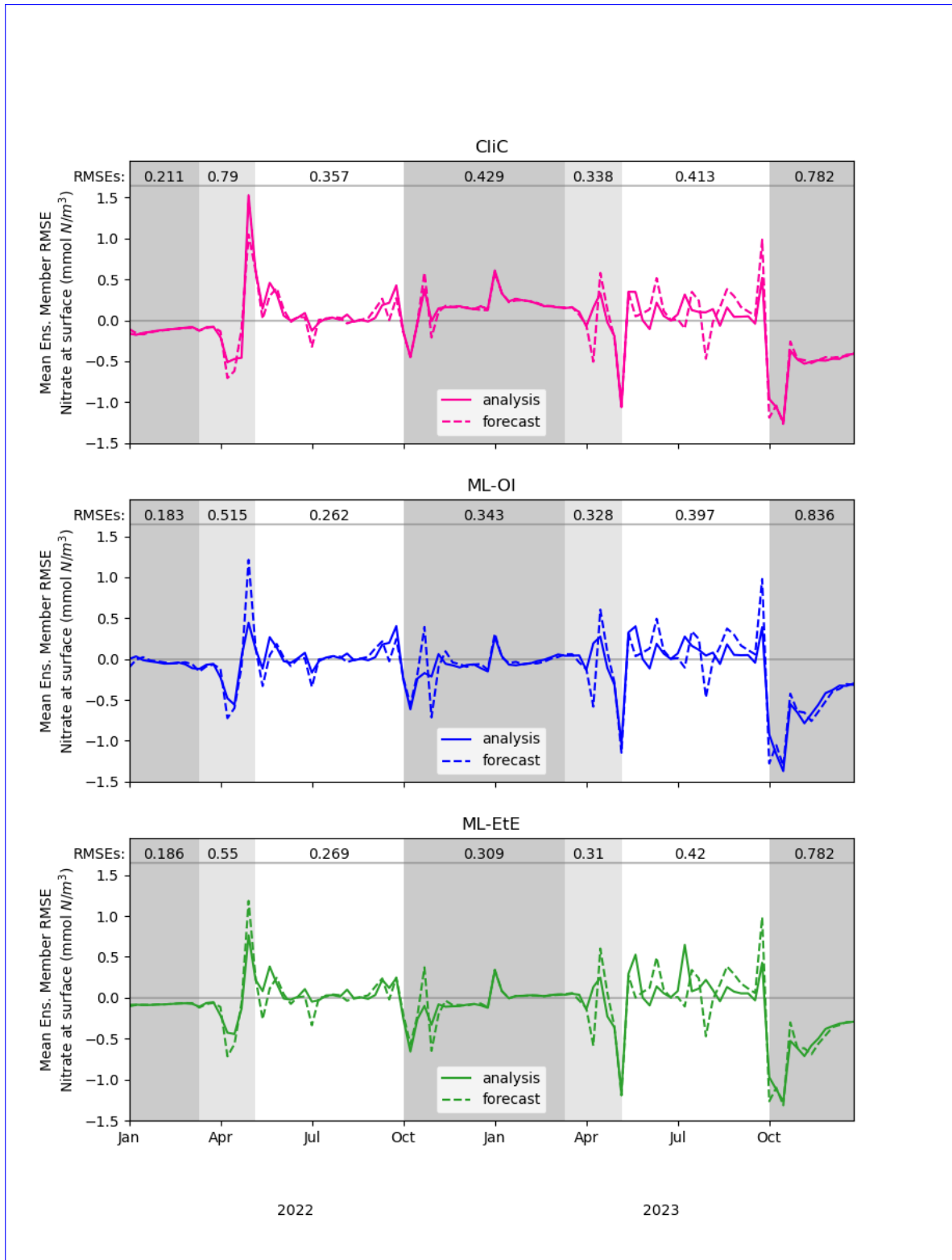


Figure 5: A comparison of mean nitrate background and analysis errors produced in the single-model runs for schemes using “online” cycled-DA at the surface. The first panel shows the analysis RMSE (solid red) and the background RMSE (dashed red) of the CliC runs. The second panel shows the analysis RMSE (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).

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

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

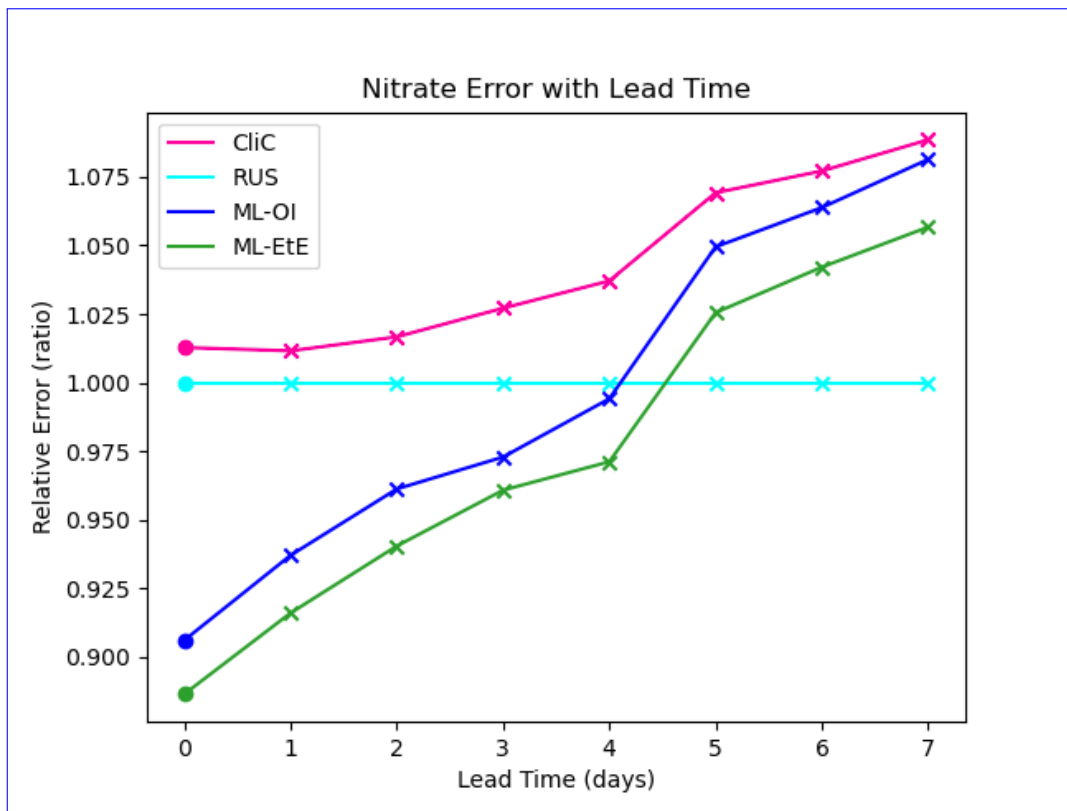


Figure 6: The forecast RMSE of each scheme relative to that of the RUS scheme at daily lead time intervals at the surface. For each scheme, the dot indicates the relative analysis error, while the crosses shows the relative forecast error for each day of lead time until a maximum lead time of 7 days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

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

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

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

662 **4.3 Extending the set of updated variables**

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

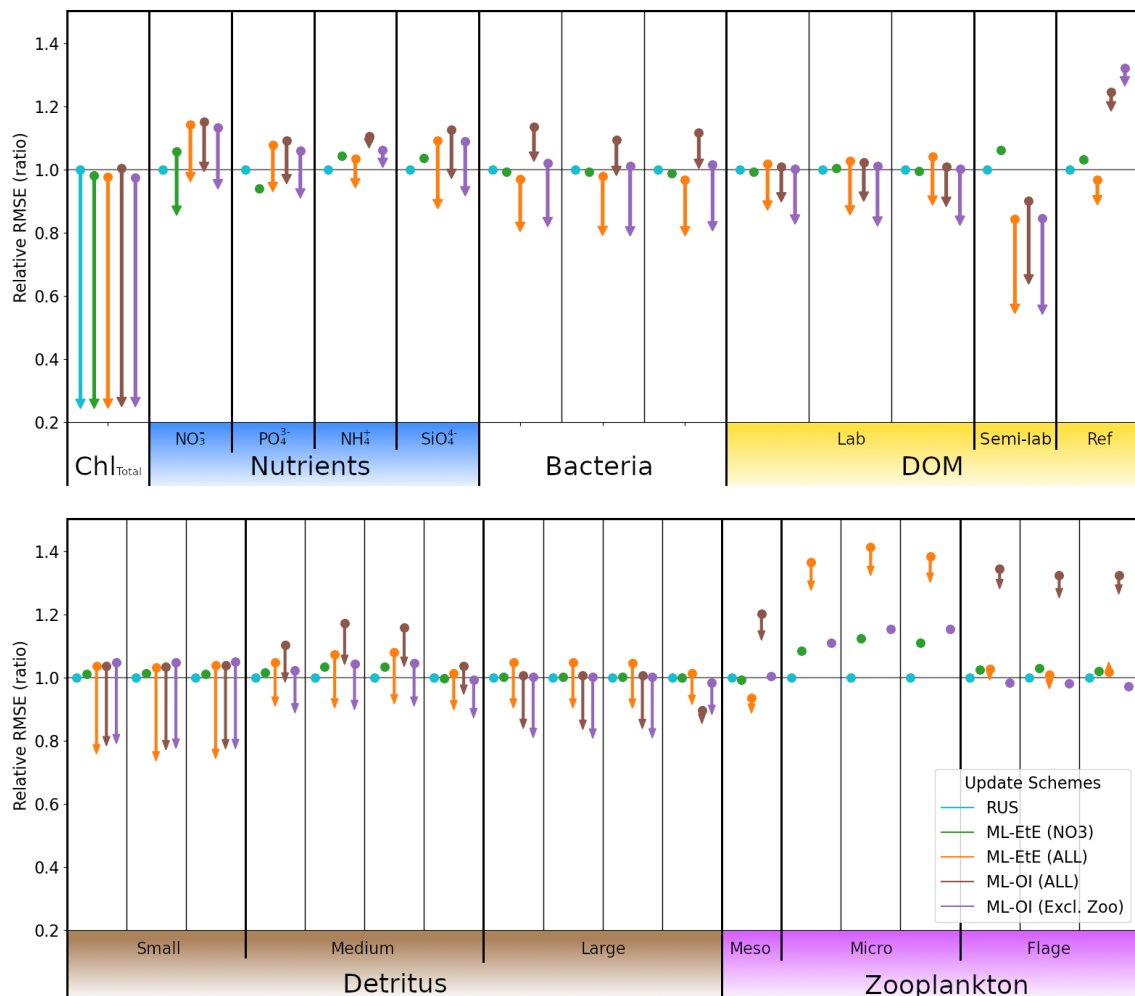


Figure 7: A comparison of the different schemes implemented to update the various components of the ERSEM BGC model at the L4 location. Surface RMSEs are calculated as an average across the entire online test period and across all initialisations. Forecast state (and equivalently background state) RMSEs 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 (NO₃) 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.

667 The RUS scheme, described in Sect. 2.4.1, is used as a benchmark for the extended schemes, and
 668 so values shown in Fig. 7 are RMSEs for 7-day forecasts relative to the RMSE of the RUS method
 669 (averaged across the 104 forecast-analysis cycles of the entire test period). Again, we recall that the
 670 RUS does not update any variables beyond total chlorophyll (shown) and its constituent PFTs (not
 671 shown). The ML-EtE (NO₃) scheme (green), which updates only nitrate, is carried over from the
 672 previous section (as it performed best), to act as another point of comparison for the extended schemes.
 673 Before discussing the extended schemes, we can see from Fig. 7 the dynamical impact that the updates
 674 of ML-EtE (NO₃) have on other (i.e. non-updated) marine BGC variables in the system. Generally,
 675 the change in RMSE for these non-updated variables is very small, with the largest improvement being
 676 to phosphate and the largest degradation to zooplankton types – particularly microzooplankton. It
 677 also slightly ~~degrades the~~ worsens the forecast error for ammonium and silicate concentrations that
 678 are not updated during the analysis. While this shows that updating a key nutrient, such as nitrate,
 679 can have wider impact on the system through dynamical adjustment, the generally beneficial results of
 680 the extended schemes (discussed below) point towards needing a DA system that can make reasonable

681 adjustments to a wider set of marine BGC variables.

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

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

707 Figure. A.2 shows that correlations between surface total chlorophyll and silicate (or phosphate) are
708 as strong as those with nitrate. This suggests that updates of similar magnitude might be expected
709 for the other nutrients as well. However, it is not straightforward to infer how the system would
710 evolve when all nutrients are updated simultaneously, based solely on the behaviour observed when
711 only nitrate is updated. For example, [FigureFig. 7](#) reveals clear differences in the forecast and analysis
712 RMSE for nitrate between the two ML-EtE configurations: one updating only nitrate and the other
713 updating all nutrients. In the case for ML-EtE (ALL), updating all nutrients degrades the nitrate
714 RMSE in [both the](#) forecast and provides negligible impact on the analysis, while improving the
715 analysis RMSEs for phosphate, ammonium and silicate. This outcome is notable, as it may point to
716 assimilation biases introduced by the choice of update strategy.

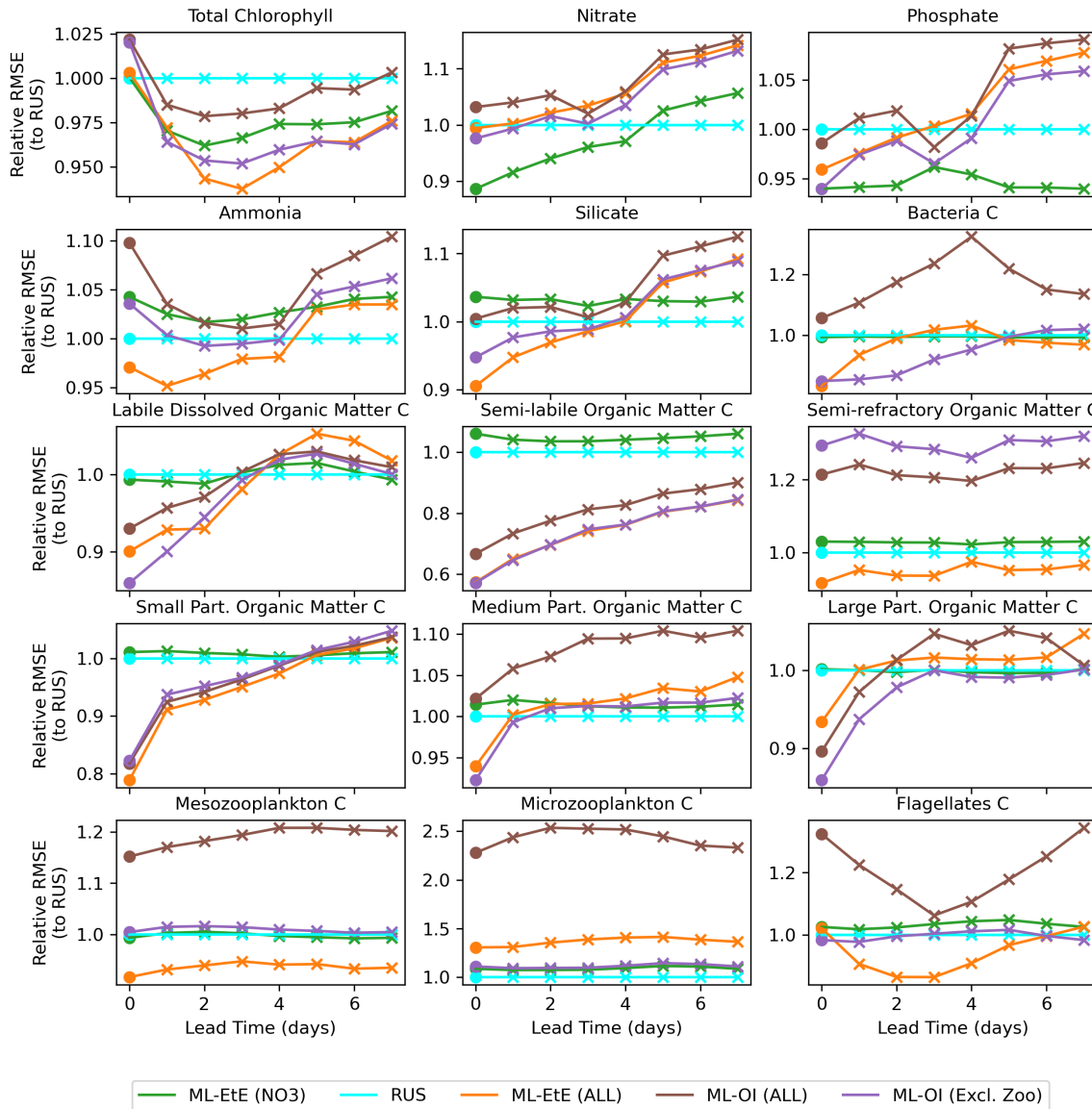


Figure 8: A comparison of the different schemes implemented to update the various components of the ERSEM BGC model at the L4 location (colours used are identical to Fig. 7). The forecast RMSE of each scheme are relative to that of the RUS scheme at daily lead time intervals. ~~the~~ The dots indicate 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 total-time between observations of total chlorophyll.

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

732 In summary the experiments from Figs. 7 and 8 show that the impact of DA analysis updates
733 on the model forecast is not straightforward due to the non-linear and complex nature of the BGC
734 model. Although in general it is true that increasing the number of updated variables benefits the
735 forecasts, especially at shorter lead times (e.g., less than 5 days), this is definitely not true for every
736 variable. Some variables, for instance, show similar or worse forecast RMSE at a 7-day lead time
737 than the RUS, which highlights the need ~~for specific results on each update strategy~~ to evaluate how
738 different assimilation strategies and variable-update choices affect performance.

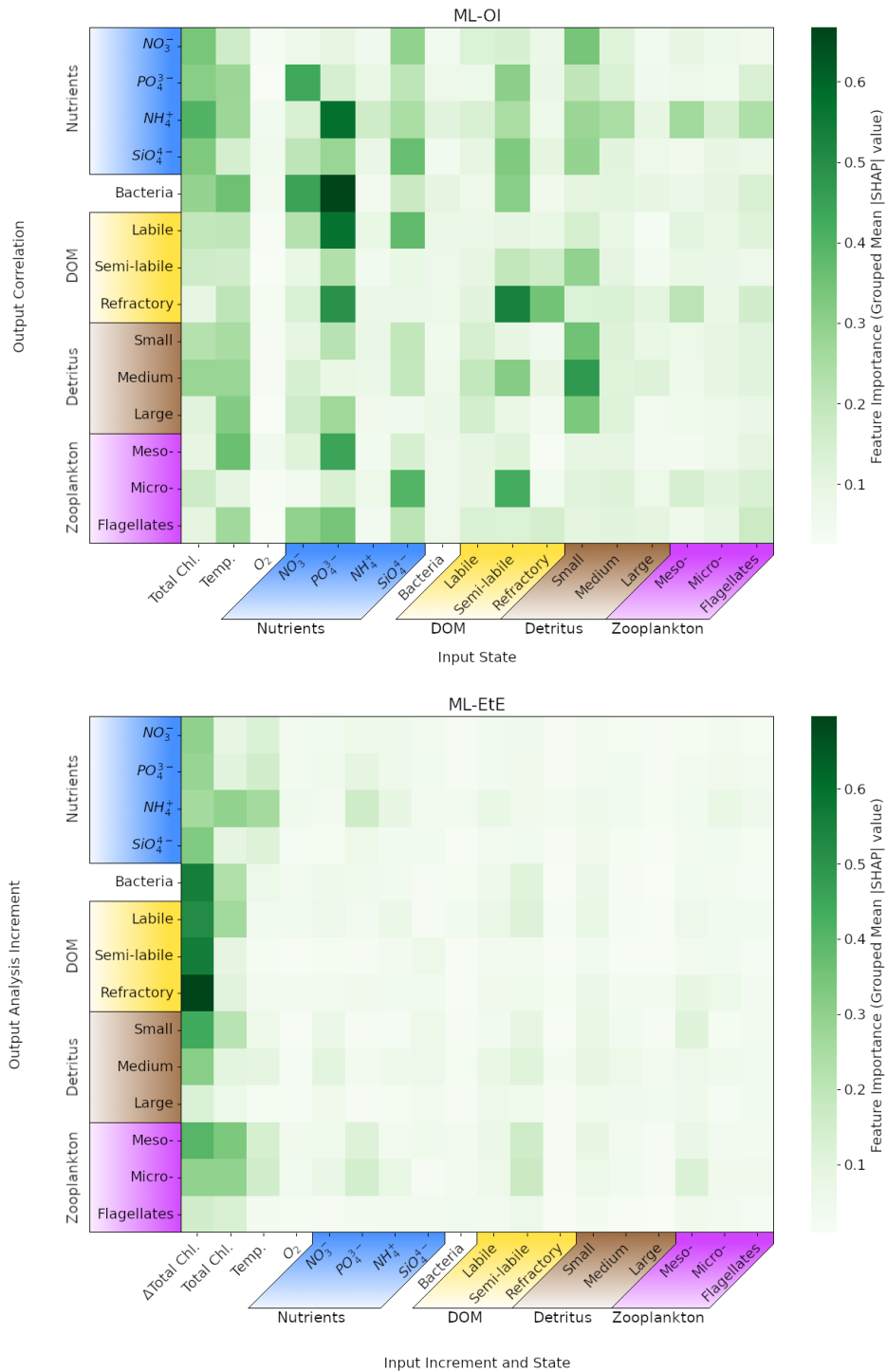


Figure 9: Mean absolute Shapley values estimated for the ML-OI (ALL) and ML-EtE (ALL), across their respective training datasets (see Sect. 3). Chemical components that belong to the same class or type (e.g., the carbon, nitrogen and phosphorus of bacteria) have been grouped as they are highly correlated. 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, Δ 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).

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

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

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

750 The ML-OI (ALL) Shapley values indicate that a broad range of input variables are important to
751 the estimation of total chlorophyll correlations with unobserved variables, and highlight the general
752 complexity of these interactions. We see that the state of temperature and total chlorophyll are moder-
753 ately important across a broad set of variable groups. This makes sense as this ML model is estimating
754 the correlation of a given variable with total chlorophyll, which is generally dependent on the state of
755 total chlorophyll. However, this also implies that the seasonal regimes play a significant role in the esti-
756 mations, as temperature is a clear identifier for the current time in the seasonal cycle. We note that the
757 seasonal signal of other variables could also be important for the estimation of correlations as, in some
758 cases, we see that at least one of the state variables in a group can be important to estimating the corre-
759 lation between total chlorophyll and a state variable of the same group. For example, the state of small
760 detritus is highly important to the entire group of detritus correlations, the states of some nutrients are
761 generally important to the estimation of nutrients, and the semi-labile DOM is somewhat important
762 to the wider DOM correlations. Each of the nutrients show moderate to strong importance across a
763 wide variety of outputs, with the strongest being phosphates. The elevated Shapley importance of
764 phosphate likely reflects interdependencies among nutrient variables, particularly given their correlated
765 seasonal behaviour, highlighting a limitation of the Shapley framework when applied to features with
766 shared temporal dynamics. Some input features show no strong importance to any output targets.
767 In particular, the zooplankton types seem largely unimportant in estimation their own correlation
768 with total chlorophyll and the variables with a stronger signal have no obvious direct relationship.
769 This may partially explain why zooplankton performs poorly when they are updated by the ML-DA
770 schemes, as seen previously in Fig. 7, and points towards the difficulty and uncertainty associated
771 with zooplankton in marine BGC modelling. This is further evidenced as the zooplankton types are
772 unimportant as input features for all other correlation estimations as well. We also see that oxygen
773 is largely unimportant to the estimation of the correlations. This observation is consistent with the
774 known weak impact of oxygen assimilation in ERSEM on other modelled variables (Skákala et al.,
775 2021). This would imply that both zooplankton and oxygen could be removed from the input feature
776 set with little impact on the overall model performance.

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

792 Given the dominant role of the total chlorophyll analysis increment in ML-EtE (ALL), it is
793 important to note that the influence of observations could differ when using real observations rather
794 than synthetic ones, owing to potential differences in error characteristics, representativeness, or
795 systematic biases. We would still expect the analysis increment of the observed variable to exhibit
796 comparable importance if trained on increments derived from real observations, as demonstrated in
797 this experiment. However, because of the strong influence of total chlorophyll increments, the overall
798 structure and nature of the resulting updates may change, reflecting the impact of more complex and
799 realistic observational error structures.

4.4 Generalisation of machine learned-correlations to an unseen location

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

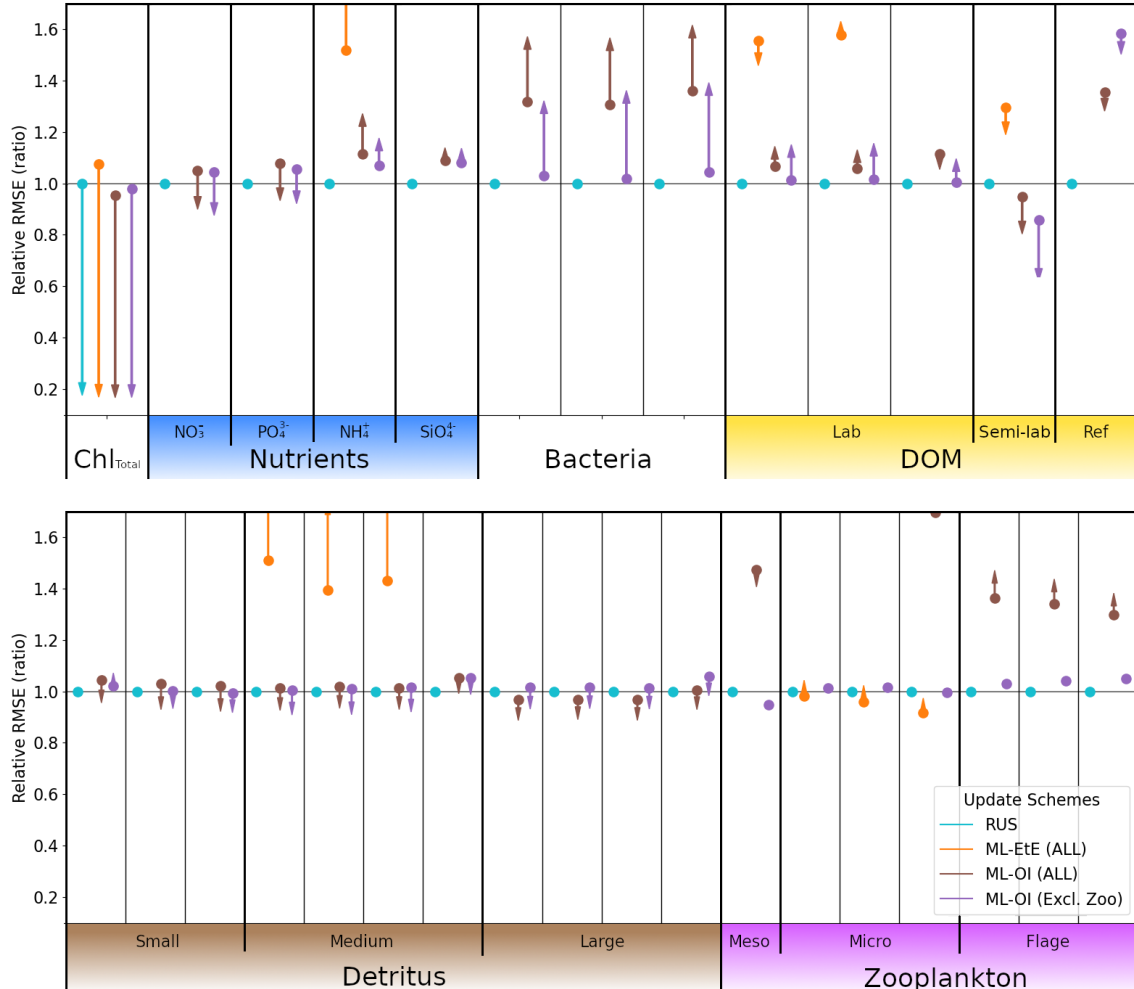


Figure 10: As Fig. 7, with the ML methods trained on L4, but applied to the new location CWEC. Dots and arrows that do not appear are off the scale.

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

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

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

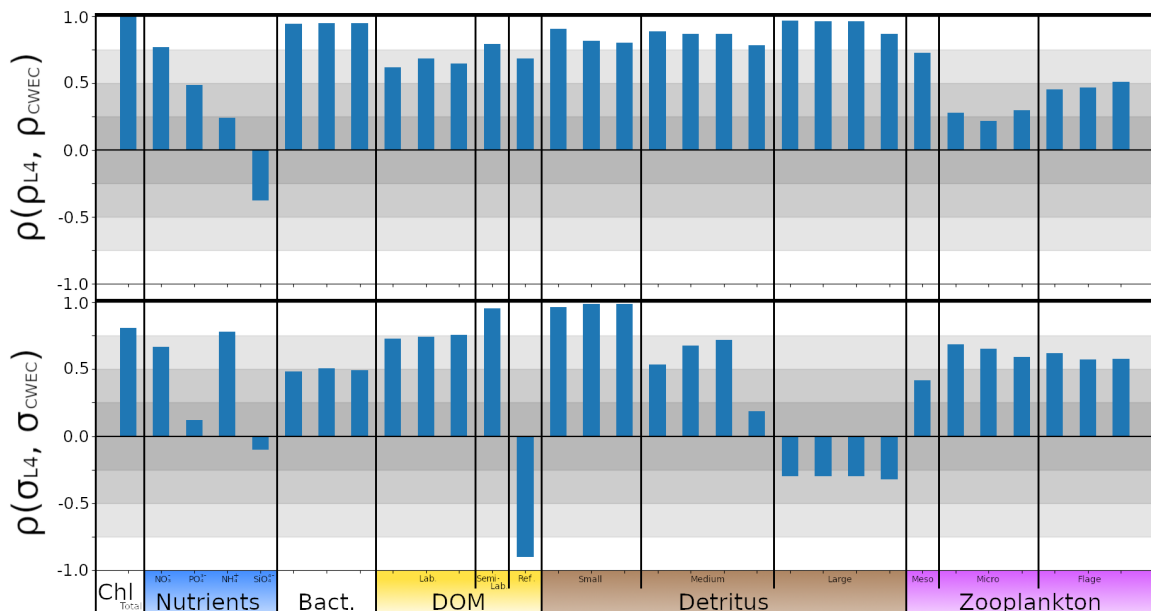


Figure 11: The top panel shows $\rho(\rho_{L4}, \rho_{CWEC})$, the correlation between a given variable's the daily climatological correlation signal correlations of each variable with total chlorophyll at L4 and at the CWEC, $\rho(\rho_{L4}, \rho_{CWEC})$. The bottom panel shows the correlation between 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.

842 4.5 Viewpoint on scaling multivariate data assimilation to 3D models

843 We have shown that ML methods can make improvements to the DA schemes of marine BGC models
 844 when coupled to a 1D physical model ~~particularly in the~~ ~~particularly shown in improved forecasts~~
 845 ~~for~~ shorter lead times ~~of (< 4 days) at~~ the training location. The natural next question is how these
 846 results would scale when the marine BGC model is coupled to a 3D physical model, such as NEMO
 847 (Nucleus for European Modelling of the Ocean). A seemingly simple solution would be to run (once)
 848 a well-tuned, large EnKF, which can then be used to train an ML model to be used operationally in
 849 an analogous 3D DA system that presently updates only total chlorophyll. A reanalysis product with
 850 comprehensive statistics, or all ensemble members available so statistics can be generated, would be
 851 ideal (Bonavita and Laloyaux, 2020; Brajard et al., 2021; Gregory et al., 2024). This circumvents the
 852 need to run an expensive DA scheme operationally as the ML model could be trained offline, and then
 853 run significantly faster while retaining the benefit of the statistics learned from a large ensemble. This

854 would also allow the analysis increments to be estimated directly. However, state-of-the-art ensemble
855 marine BGC systems are still limited in scale and may not (yet) accurately represent the statistics
856 needed for multivariate DA (Skákala et al., 2024). Also, this approach would need to be repeated
857 if/when the observation network changes, which is likely given new observation missions and strategies
858 (Telszewski et al., 2018). A cheaper alternative would be to calculate the correlations in a free-run
859 ensemble dataset, as per the methods described in Sect. 3.1. This would be cheaper to create as there
860 would be no need to store and calculate both background ~~states~~ and analysis states. However, this
861 approach cannot calculate the analysis increments directly and instead must rely on the hybridisation
862 of background covariances/correlations into existing DA frameworks. Nevertheless our results on the
863 1D scenario suggests that this is feasible and a good alternative to estimating the increments directly.

864 It is also worth considering how the data for these ML models should be sampled spatially in the
865 3D case. ~~Our results show there is some transferability between locations, as long as the dynamics
866 are similar enough.~~ In this, we suggest that a sparse forest of 1D models could be generated across
867 the 3D domain, which aims to cover each region of sufficiently different biogeochemical behaviour.
868 Previous work by Higgs et al. (2024) has split the North-West European Shelf into dynamically
869 connected ecoregions, and this, or similar analysis, could be used as a guideline for generating these
870 1D models. Furthermore, ML models could handle local multivariate aspects (a 0D transformation),
871 while traditional DA methods (such as spatial correlations functions) manage 3D reconstruction (just
872 as they manage the 1D reconstruction in our setup). A limitation of our two test locations is that
873 they are not directly coupled, and could only be considered weakly coupled in the sense that their
874 forcing data is extracted from the same 3D weather model. This could mean that 3D models have an
875 advantage in locations having similar behaviour, as model grid points are much more likely to strongly
876 correlate due to advection and ocean currents. However, the inverse could also be true, as the 1D
877 models do not consider riverine input which can have substantial effects at the coast. Either way, the
878 results suggest that some sparsity could be applied in extracting training data for these models, as
879 long as each regime of BGC behaviour is represented in the selection. Introducing spatial variables
880 like longitude and latitude could also improve the models ability to estimate increments or correlations
881 across the different horizontal locations.

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

892 5 Conclusions

893 Marine biogeochemistry (BGC) models aim to represent the complex BGC processes necessary to
894 understand and forecast ecosystem behaviour. Data assimilation (DA) plays a crucial role in ensur-
895 ing model trajectories remain closely aligned with real-world observations, along with the need for
896 continuous improvement of numerical estimations. In this study, we used a synthetic “perfect model”
897 setup as a ~~necessary~~ first step to explore ML-assisted DA under controlled conditions, but a natural
898 direction for future work is to apply these approaches to real observations to assess their practical
899 value. ~~Nevertheless, both~~ The ML-based methods considered in this study extend a conventional
900 univariate DA system (RUS, see Table 2), which observes only total chlorophyll. One ML method
901 (ML-OI) estimates the flow-dependent background error correlations between total chlorophyll and
902 other model variables (which are then used in an otherwise conventional DA update step). The other
903 ML method (ML-EtE) directly estimates the analysis increments of other model variables (bypassing
904 the conventional DA update equations for these variables).

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

911 in italics).

912 *(a) Can we make improvements to the existing univariate scheme by updating a limited set of*
913 *additional variables with an ML model to estimate correlations or analysis increments?* In this study,
914 we have demonstrated that neural networks can effectively learn statistical relationships between
915 total chlorophyll (the only observed variable) and various pelagic BGC model variables. With ma-
916 chine learning (ML), we achieve improvements over climatological statistics in the nitrate-only update
917 framework. Our analysis of ML-estimated nitrate updates illustrates that the ML methods behave in
918 a largely coherent and meaningful manner. While ML can degrade forecast skill in some unobserved
919 variables compared to RUS or nitrate-only ML schemes, they nonetheless show promise, with their
920 usefulness in more complex assimilation settings requiring further assessment.

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

933 *(c) Is the ML model transferable to a new location after being trained on some other location?*

934 While a neural network trained in one water column exhibits partial transferability to other locations,
935 challenges remain in fully generalising the model across spatial domains. This partial transferability
936 is valuable, given the difficulty and cost of acquiring high-quality training data across large oceanic
937 regions, and should be explored further in the context of 3D models. We discuss the feasibility of
938 this, and propose a methodology for doing so. Future work should focus on refining transferability
939 strategies, effective sampling strategy to allow for ergodic coverage (i.e., ensuring statistical represen-
940 tativeness over space and/or time), and further evaluating the scalability of ML-driven DA in complex
941 marine environments.

942 A Characterisation of location biogeochemistry

943 Figure A.1 shows ~~that the CEWC~~ the variability of each ERSEM variable over the online testing period
 944 for the L4 and CWEC locations. It shows that the CWEC is clearly less biologically productive than
 945 L4 with surface concentrations of total chlorophyll having a significantly lower median value, and a
 946 maximum that is approximately 50% of L4's maximum. Each ~~exhibit~~ exhibits similar temperature
 947 values, as they are both located within the English Channel. In the nutrients group, nitrate and
 948 phosphate values cover a similar range in each location, but ammonium and silicate have little overlap.
 949 Bacteria and DOM concentrations also show little similarity between locations. The small detritus
 950 concentrations are very similar between both locations, but the medium and large detritus differ
 951 significantly, with CWEC covering a much wider range of values than L4. Zooplankton concentrations
 952 also differ between the locations, with CWEC producing much lower concentrations of zooplankton
 953 than the more biologically active L4 location.

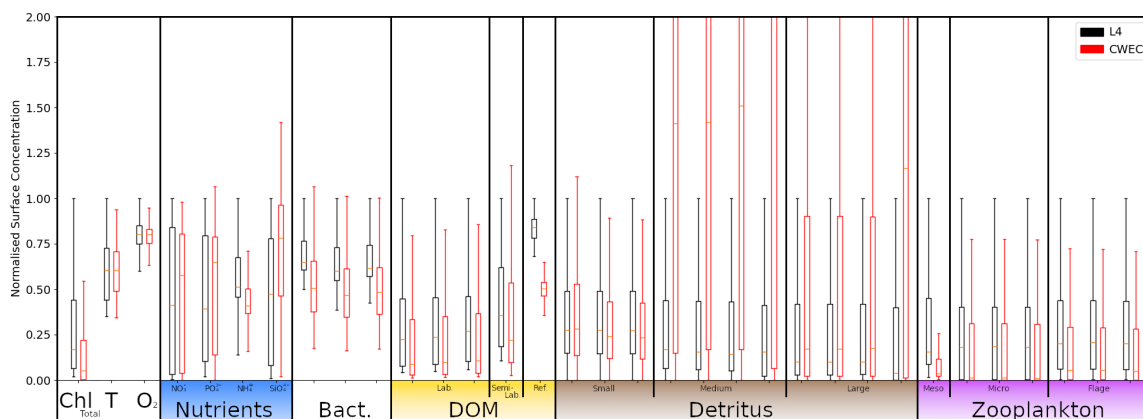


Figure A.1: Box and whisker plot showing the 25th, 50th and 75th percentile and upper and lower bound (excluding outliers larger than $1.5 \times$ Inter quartile range) of each pelagic marine BGC variable for the RUS scheme in the online testing period. Values for L4 are given in black, and values for 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.

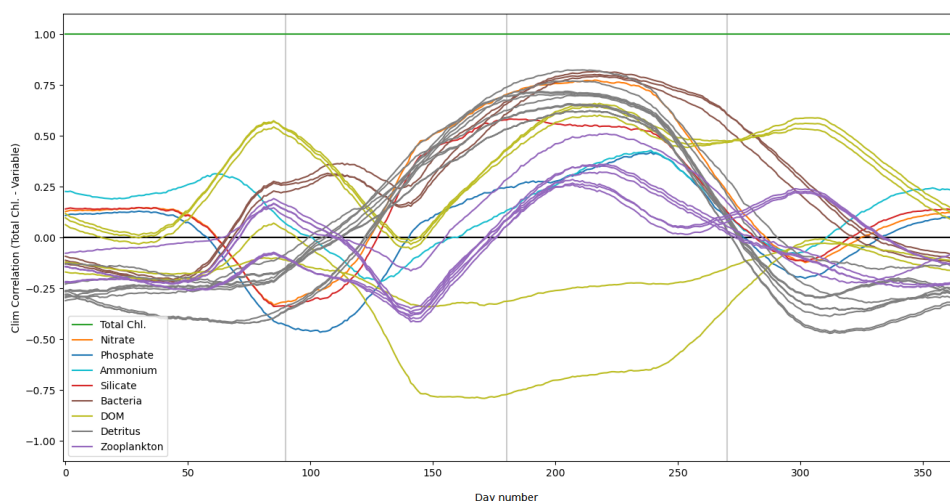


Figure A.2: Daily climatological correlations ~~for~~ between total chlorophyll and each pelagic variable at the L4 location, calculated from the training free-run period of 2000-2014. Pelagic variables of the same class (according to Table 2) are shown with the same colour, except nutrients (nitrate,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

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

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

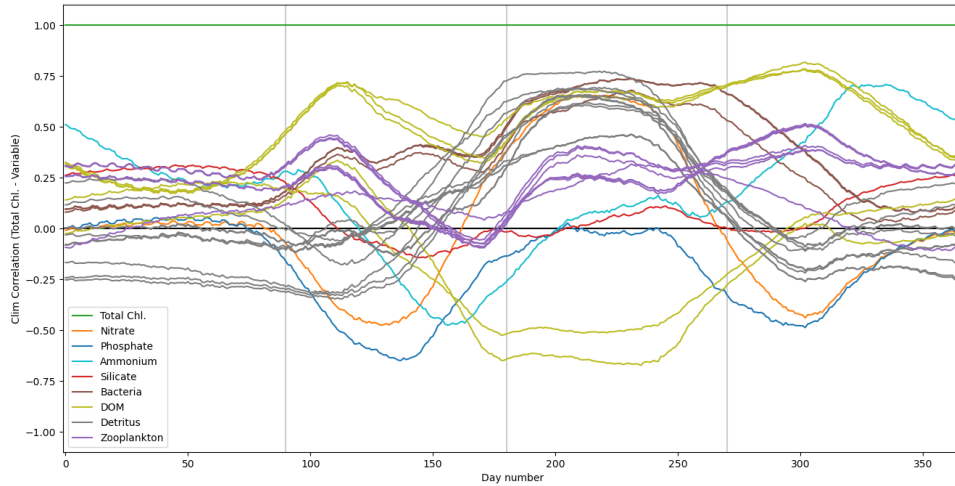


Figure A.3: As with Fig.A.2, except for the CWEC location from a free-run period of 2000-2010.

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

965 B Dynamical impact of updating only nitrate

966 Figure B.1 [shows is](#) an extension of Fig. 4, with a representative set of variables that are unobserved and
 967 not updated (unlike nitrate which is also unobserved, but updated in this experiment). This clearly
 968 shows that the improvement of nitrate does not necessarily translate to an improvement in other
 969 variables, regardless of the method used to update the nitrate. This highlights the need [for specific](#)
 970 [results of each update strategy to evaluate how different assimilation strategies and variable-update](#)
 971 [choices affect performance.](#)

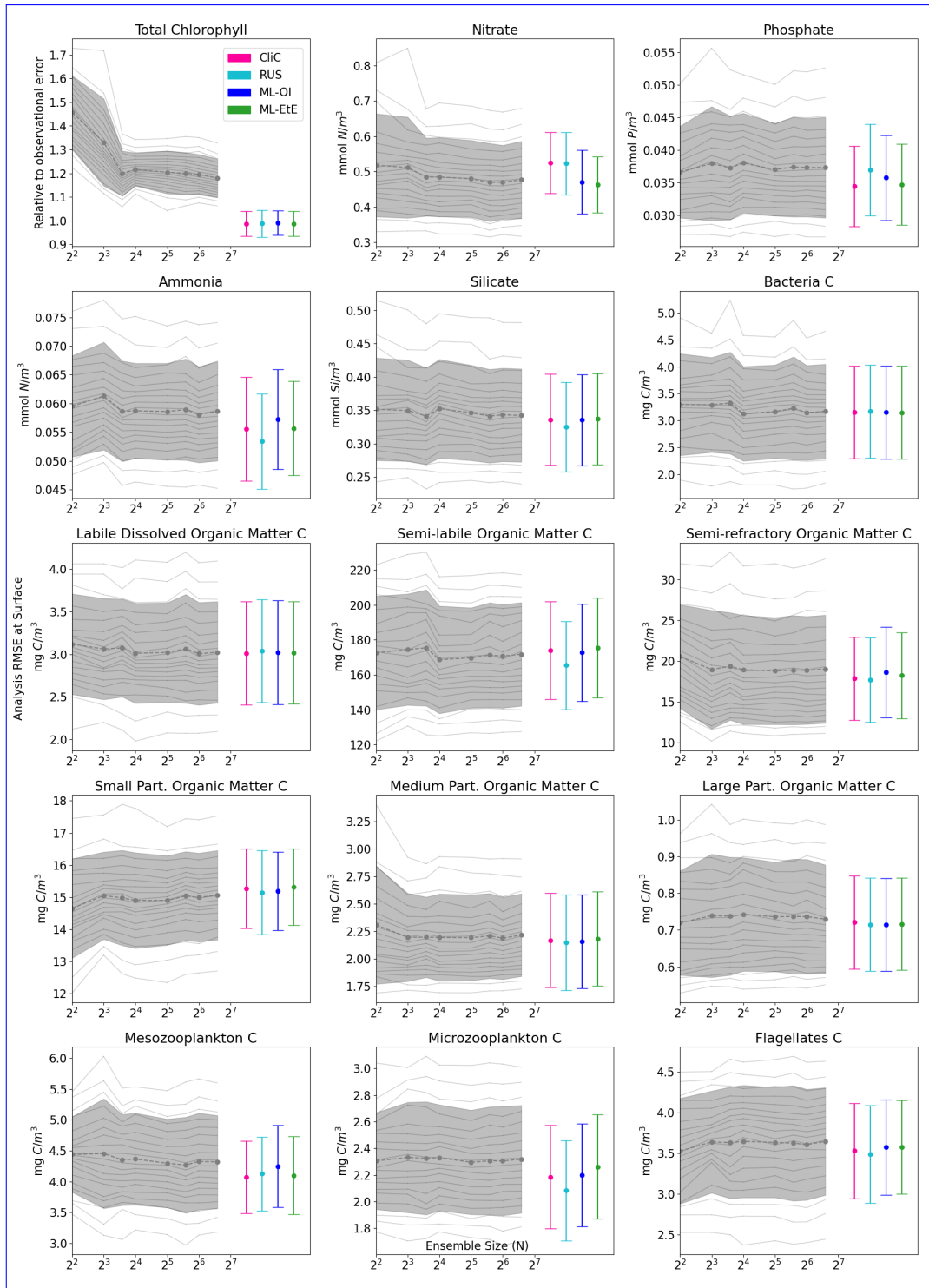


Figure B.1: An extension of Fig. 4, where panels for total chlorophyll (observed) and nitrate (un-observed, but updated) are the same. **A**-representative set of additional variables (unobserved, not updated) are shown. A “C” indicates the panel is given for the carbon chemical component in the model.

972 Author Contributions

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

975 **Competing interests**

976 ~~At least one~~ One of the (co-)authors is a member of the editorial board of Biogeosciences. The peer-
977 review process was guided by an independent editor, and the authors also have no other competing
978 interests to declare.

References

- 979
- 980 Anugerahanti, P., Kerimoglu, O., and Smith, S. L. (2021). Enhancing ocean biogeochemical models
981 with phytoplankton variable composition. *Frontiers in Marine Science*, 8:944.
- 982 Artioli, Y., Blackford, J. C., Butenschön, M., Holt, J. T., Wakelin, S. L., Thomas, H., Borges, A. V.,
983 and Allen, J. I. (2012). The carbonate system in the North Sea: Sensitivity and model validation.
984 *Journal of Marine Systems*, 102:1–13.
- 985 Asch, M., Bocquet, M., and Nodet, M. (2016). *Data assimilation: methods, algorithms, and applica-*
986 *tions*. SIAM.
- 987 Baretta, J., Ebenhöf, W., and Ruardij, P. (1995). The European regional seas ecosystem model, a
988 complex marine ecosystem model. *Netherlands Journal of Sea Research*, 33(3-4):233–246.
- 989 Baretta-Bekker, J., Baretta, J., and Ebenhöf, W. (1997). Microbial dynamics in the marine ecosystem
990 model ERSEM II with decoupled carbon assimilation and nutrient uptake. *Journal of Sea Research*,
991 38(3-4):195–211.
- 992 Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M. (2020). DINCAE 1.0: A convolutional
993 neural network with error estimates to reconstruct sea surface temperature satellite observations.
994 *Geoscientific Model Development*, 13(3):1609–1622.
- 995 Bertino, L., Ali, A., Carrasco, A., Lien, V., and Melsom, A. (2021). The Arctic Marine Forecasting
996 Center in the first Copernicus period. In *9th EuroGOOS International conference*, pages 256–263.
- 997 Blackford, J. (1997). An analysis of benthic biological dynamics in a North Sea ecosystem model.
998 *Journal of Sea Research*, 38(3-4):213–230.
- 999 Bocquet, M., Farchi, A., Finn, T. S., Durand, C., Cheng, S., Chen, Y., Pasmans, I., and Carrassi,
1000 A. (2024). Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities
1001 without an ensemble. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(9).
- 1002 Bolding, K. and Villarreal, M. R. (1999). GOTM: A general ocean turbulence model: Theory, appli-
1003 cations and test cases. Technical report, European Commission Tech. Rep. EUR 18745 EN.
- 1004 Bonavita, M. and Laloyaux, P. (2020). Machine learning for model error inference and correction.
1005 *Journal of Advances in Modeling Earth Systems*, 12(12):e2020MS002232.
- 1006 Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2020). Combining data assimilation and
1007 machine learning to emulate a dynamical model from sparse and noisy observations: A case study
1008 with the Lorenz 96 model. *Journal of computational science*, 44:101171.
- 1009 Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L. (2021). Combining data assimilation and
1010 machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal*
1011 *Society A*, 379(2194):20200086.
- 1012 Bruggeman, J. and Bolding, K. (2014). A general framework for aquatic biogeochemical models.
1013 *Environmental modelling & software*, 61:249–265.
- 1014 Bruggeman, J., Bolding, K., Nerger, L., Teruzzi, A., Spada, S., Skakala, J., Ciavatta, S., et al.
1015 (2024). Eat v1. 0.0: a 1d test bed for physical–biogeochemical data assimilation in natural waters.
1016 *GEOSCIENTIFIC MODEL DEVELOPMENT*, 17(14):5619–5639.
- 1017 Buizza, C., Casas, C. Q., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur,
1018 T., Ruiz, L. B., et al. (2022). Data learning: Integrating data assimilation and machine learning.
1019 *Journal of Computational Science*, 58:101525.
- 1020 Burchard, H. (2009). Combined effects of wind, tide, and horizontal density gradients on stratification
1021 in estuaries and coastal seas. *Journal of Physical Oceanography*, 39(9):2117–2136.
- 1022 Butenschön, M., Clark, J., Aldridge, J. N., Allen, J. I., Artioli, Y., Blackford, J., Bruggeman, J.,
1023 Cazenave, P., Ciavatta, S., Kay, S., et al. (2016). ERSEM 15.06: a generic model for marine biogeo-
1024 chemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development*,
1025 9(4):1293–1339.

- 1026 Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences:
1027 An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*,
1028 9(5):e535.
- 1029 Cheng, S., Quilodr an-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss,
1030 B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., and Arcucci,
1031 R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical
1032 systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387.
- 1033 Ciavatta, S., Brewin, R., Skakala, J., Polimene, L., de Mora, L., Artioli, Y., and Allen, J. I. (2018).
1034 Assimilation of ocean-color plankton functional types to improve marine ecosystem simulations.
1035 *Journal of Geophysical Research: Oceans*, 123(2):834–854.
- 1036 Ciavatta, S., Kay, S., Brewin, R. J., Cox, R., Di Cicco, A., Nencioli, F., Polimene, L., Sammartino, M.,
1037 Santoleri, R., Skakala, J., et al. (2019). Ecoregions in the Mediterranean Sea through the reanalysis
1038 of phytoplankton functional types and carbon fluxes. *Journal of Geophysical Research: Oceans*,
1039 124(10):6737–6759.
- 1040 Ciavatta, S., Kay, S., Saux-Picart, S., Butensch on, M., and Allen, J. (2016). Decadal reanalysis of
1041 biogeochemical indicators and fluxes in the North West European shelf-sea ecosystem. *Journal of*
1042 *Geophysical Research: Oceans*, 121(3):1824–1845.
- 1043 Ciavatta, S., Torres, R., Martinez-Vicente, V., Smyth, T., Dall’Olmo, G., Polimene, L., and Allen,
1044 J. I. (2014). Assimilation of remotely-sensed optical properties to improve marine biogeochemistry
1045 modelling. *Progress in Oceanography*, 127:74–95.
- 1046 Ciliberti, S. A., Gr egoire, M., Staneva, J., Palazov, A., Coppini, G., Lecci, R., Peneva, E., Matreata,
1047 M., Marinova, V., Masina, S., et al. (2021). Monitoring and forecasting the ocean state and bio-
1048 geochemical processes in the Black Sea: Recent developments in the Copernicus Marine Service.
1049 *Journal of Marine Science and Engineering*, 9(10):1146.
- 1050 Coppini, G., Clementi, E., Cossarini, G., Korres, G., Drudi, M., Amadio, C., Aydogdu, A., Agostini,
1051 P., Bolzon, G., Cret i, S., et al. (2021). The Copernicus marine service ocean forecasting system for
1052 the Mediterranean Sea. In *9th EuroGOOS International conference*, pages 272–279.
- 1053 Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., and
1054 d’Ortenzio, F. (2019). Towards operational 3D-Var assimilation of chlorophyll biogeochemical-Argo
1055 float data into a biogeochemical model of the Mediterranean Sea. *Ocean Modelling*, 133:112–128.
- 1056 Cossarini, G., Querin, S., Solidoro, C., Sannino, G., Lazzari, P., Di Biagio, V., and Bolzon, G. (2017).
1057 Development of BFMCOUPLER (v1. 0), the coupling scheme that links the MITgcm and BFM
1058 models for ocean biogeochemistry simulations. *Geoscientific Model Development*, 10(4):1423–1445.
- 1059 Council, N. R., on Geosciences, C., Science, W., Board, T., Board, O. S., on the Causes, C., and
1060 of Coastal Eutrophication, M. (2000). *Clean Coastal Waters: Understanding and Reducing the*
1061 *Effects of Nutrient Pollution*. National Academies Press.
- 1062 Doney, S. C., Fabry, V. J., Feely, R. A., and Kleypas, J. A. (2009). Ocean acidification: the other
1063 CO2 problem. *Annual review of marine science*, 1(1):169–192.
- 1064 Dowd, M., Jones, E., and Parslow, J. (2014). A statistical overview and perspectives on data assimi-
1065 lation for marine biogeochemical models. *Environmetrics*, 25(4):203–213.
- 1066 Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementa-
1067 tion. *Ocean dynamics*, 53:343–367.
- 1068 Fablet, R., Chapron, B., Drumetz, L., M emin, E., Pannekoucke, O., and Rousseau, F. (2021). Learning
1069 variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*,
1070 13(10):e2021MS002572.
- 1071 Falchetti, S., Conley, D. C., Brocchini, M., and Elgar, S. (2010). Nearshore bar migration and
1072 sediment-induced buoyancy effects. *Continental Shelf Research*, 30(2):226–238.
- 1073 Fennel, K., Gehlen, M., Brasseur, P., Brown, C. W., Ciavatta, S., Cossarini, G., Crise, A., Edwards,
1074 C. A., Ford, D., Friedrichs, M. A., et al. (2019). Advancing marine biogeochemical and ecosystem
1075 reanalyses and forecasts as tools for monitoring and managing ecosystem health. *Frontiers in Marine*
1076 *Science*, 6:89.

- 1077 Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., and Yu, L. (2022). Ocean
1078 biogeochemical modelling. *Nature Reviews Methods Primers*, 2(1):76.
- 1079 Fennel, K. and Testa, J. M. (2019). Biogeochemical controls on coastal hypoxia. *Annual Review of*
1080 *Marine Science*, 11(1):105–130.
- 1081 Ford, D., Edwards, K., Lea, D., Barciela, R., Martin, M., and Demaria, J. (2012). Assimilating
1082 GlobColour ocean colour data into a pre-operational physical-biogeochemical model. *Ocean Science*,
1083 8(5):751–771.
- 1084 Ford, D., Key, S., McEwan, R., Totterdell, I., and Gehlen, M. (2018). Marine biogeochemical modelling
1085 and data assimilation for operational forecasting, reanalysis, and climate research. *New Frontiers*
1086 *in Operational Oceanography*, pages 625–652.
- 1087 Frölicher, T. L. and Laufkötter, C. (2018). Emerging risks from marine heat waves. *Nature commu-*
1088 *nications*, 9(1):650.
- 1089 Galli, G., Wakelin, S., Harle, J., Holt, J., and Artioli, Y. (2024). Multi-model comparison of trends
1090 and controls of near-bed oxygen concentration on the Northwest European Continental Shelf under
1091 climate change. *Biogeosciences*, 21(8):2143–2158.
- 1092 Gehlen, M., Barciela, R., Bertino, L., Bresseur, P., Butenschön, M., Chai, F., Crise, A., Drillet, Y.,
1093 Ford, D., Lavoie, D., et al. (2015). Building the capacity for forecasting marine biogeochemistry
1094 and ecosystems: recent advances and future developments. *Journal of Operational Oceanography*,
1095 8(sup1):s168–s187.
- 1096 Geider, R., MacIntyre, H., and Kana, T. (1997). Dynamic model of phytoplankton growth and
1097 acclimation: responses of the balanced growth rate and the chlorophyll a: carbon ratio to light,
1098 nutrient-limitation and temperature. *Marine Ecology Progress Series*, 148:187–200.
- 1099 Gobler, C. J. (2020). Climate change and harmful algal blooms: insights and perspective. *Harmful*
1100 *algae*, 91:101731.
- 1101 Gregg, W. W. and Rousseaux, C. S. (2017). Simulating pace global ocean radiances. *Frontiers in*
1102 *Marine Science*, 4:60.
- 1103 Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., and Zanna, L. (2024). Machine learning for
1104 online sea ice bias correction within global ice-ocean simulations. *Geophysical Research Letters*,
1105 51(3):e2023GL106776.
- 1106 Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., Brockmann, C.,
1107 Chauhan, P., Choi, J.-k., Chuprin, A., et al. (2019). Satellite ocean colour: Current status and
1108 future perspective. *Frontiers in Marine Science*, 6:485.
- 1109 Gutknecht, E., Refray, G., Mignot, A., Dabrowski, T., and Sotillo, M. G. (2019). Modelling the marine
1110 ecosystem of Iberia–Biscay–Ireland (IBI) European waters for CMEMS operational applications.
1111 *Ocean Science*, 15(6):1489–1516.
- 1112 Hayashida, H., Steiner, N., Monahan, A., Galindo, V., Lizotte, M., and Lavoie, M. (2017). Impli-
1113 cations of sea-ice biogeochemistry for oceanic production and emissions of dimethyl sulfide in the
1114 Arctic. *Biogeosciences*, 14(12):3129–3155.
- 1115 Heinze, C. and Gehlen, M. (2013). Modeling ocean biogeochemical processes and the resulting tracer
1116 distributions. In *International Geophysics*, volume 103, pages 667–694. Elsevier.
- 1117 Hemmings, J. C., Barciela, R. M., and Bell, M. J. (2008). Ocean color data assimilation with material
1118 conservation for improving model estimates of air-sea CO₂ flux.
- 1119 Higgs, I., Skákala, J., Bannister, R., Carrassi, A., and Ciavatta, S. (2024). Investigating ecosystem
1120 connections in the shelf sea environment using complex networks. *Biogeosciences*, 21(3):731–746.
- 1121 Hu, Q., Zhang, R., and Zhou, Y. (2016). Transfer learning for short-term wind speed prediction with
1122 deep neural networks. *Renewable Energy*, 85:83–95.
- 1123 Jin, H., Song, Q., and Hu, X. (2019). Auto-keras: An efficient neural architecture search system.
1124 In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data*
1125 *mining*, pages 1946–1956.

- 1126 Jones, E. M., Baird, M. E., Mongin, M., Parslow, J., Skerratt, J., Lovell, J., Margvelashvili, N.,
1127 Matear, R. J., Wild-Allen, K., Robson, B., et al. (2016). Use of remote-sensing reflectance to con-
1128 strain a data assimilating marine biogeochemical model of the Great Barrier Reef. *Biogeosciences*,
1129 13(23):6441–6469.
- 1130 Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects.
1131 *Science*, 349(6245):255–260.
- 1132 Kerimoglu, O., Anugerahanti, P., and Smith, S. L. (2021). FABM-NflexPD 1.0: assessing an instantane-
1133 ous acclimation approach for modeling phytoplankton growth. *Geoscientific Model Development*,
1134 14(10):6025–6047.
- 1135 Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., and Hoyer, S. (2021). Machine
1136 learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sci-
1137 ences*, 118(21):e2101784118.
- 1138 Le Traon, P. Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., Bentamy,
1139 A., Bertino, L., Brando, V. E., Kreiner, M. B., et al. (2019). From observation to information and
1140 users: The copernicus marine service perspective. *Frontiers in Marine Science*, 6:234.
- 1141 Leeds, W., Wikle, C., Fiechter, J., Brown, J., and Milliff, R. (2013). Modeling 3-D spatio-temporal
1142 biogeochemical processes with a forest of 1-D statistical emulators. *Environmetrics*, 24(1):1–12.
- 1143 Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances
1144 in neural information processing systems*, 30.
- 1145 Mandal, S., Homma, H., Priyadarshi, A., Burchard, H., Smith, S. L., Wirtz, K. W., and Yamazaki,
1146 H. (2016). A 1D physical–biological model of the impact of highly intermittent phytoplankton
1147 distributions. *Journal of Plankton Research*, 38(4):964–976.
- 1148 Mattern, J. P., Fennel, K., and Dowd, M. (2012). Estimating time-dependent parameters for a
1149 biological ocean model using an emulator approach. *Journal of Marine Systems*, 96:32–47.
- 1150 Mattern, J. P., Fennel, K., and Dowd, M. (2013). Sensitivity and uncertainty analysis of model hypoxia
1151 estimates for the Texas-Louisiana shelf. *Journal of Geophysical Research: Oceans*, 118(3):1316–
1152 1332.
- 1153 Mattern, J. P., Fennel, K., and Dowd, M. (2014). Periodic time-dependent parameters improving
1154 forecasting abilities of biological ocean models. *Geophysical Research Letters*, 41(19):6848–6854.
- 1155 Mattern, J. P., Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J. (2017). Data assimilation of
1156 physical and chlorophyll a observations in the California Current System using two biogeochemical
1157 models. *Ocean Modelling*, 109:55–71.
- 1158 McEwan, R., Kay, S., and Ford, D. (2021). Quality information document for the CMEMS North
1159 West European Shelf biogeochemical analysis and forecast. Technical report, CMEMS-NWS-QUID-
1160 004-002 report.
- 1161 Nowack, P., Braesicke, P., Haigh, J., Abraham, N. L., Pyle, J., and Voulgarakis, A. (2018). Us-
1162 ing machine learning to build temperature-based ozone parameterizations for climate sensitivity
1163 simulations. *Environmental Research Letters*, 13(10):104016.
- 1164 Ouala, S., Fablet, R., Herzet, C., Chapron, B., Pascual, A., Collard, F., and Gaultier, L. (2018).
1165 Neural network based Kalman filters for the spatio-temporal interpolation of satellite-derived sea
1166 surface temperature. *Remote Sensing*, 10(12):1864.
- 1167 Pingree, R. and Griffiths, D. (1978). Tidal fronts on the shelf seas around the British Isles. *Journal
1168 of Geophysical Research: Oceans*, 83(C9):4615–4622.
- 1169 Pradhan, H. K., Völker, C., Losa, S. N., Bracher, A., and Nerger, L. (2020). Global assimilation
1170 of ocean-color data of phytoplankton functional types: Impact of different data sets. *Journal of
1171 Geophysical Research: Oceans*, 125(2):e2019JC015586.
- 1172 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f.
1173 (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*,
1174 566(7743):195–204.

- 1175 Sacco, M. A., Pulido, M., Ruiz, J. J., and Tandeo, P. (2024). On-line machine-learning forecast un-
 1176 certainty estimation for sequential data assimilation. *Quarterly Journal of the Royal Meteorological*
 1177 *Society*, 150(762):2937–2954.
- 1178 Sacco, M. A., Ruiz, J. J., Pulido, M., and Tandeo, P. (2022). Evaluation of machine learning tech-
 1179 niques for forecast uncertainty quantification. *Quarterly Journal of the Royal Meteorological Society*,
 1180 148(749):3470–3490.
- 1181 Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., Ward, B. A., Slawig,
 1182 T., and Oschlies, A. (2017). Reviews and syntheses: parameter identification in marine planktonic
 1183 ecosystem modelling. *Biogeosciences*, 14(6):1647–1701.
- 1184 Schmidtke, S., Stramma, L., and Visbeck, M. (2017). Decline in global oceanic oxygen content during
 1185 the past five decades. *Nature*, 542(7641):335–339.
- 1186 Shulman, I., Frolov, S., Anderson, S., Penta, B., Gould, R., Sakalaukus, P., and Ladner, S. (2013). Im-
 1187 pact of bio-optical data assimilation on short-term coupled physical, bio-optical model predictions.
 1188 *Journal of Geophysical Research: Oceans*, 118(4):2215–2230.
- 1189 Simon, E. and Bertino, L. (2012). Gaussian anamorphosis extension of the DEKF for combined state
 1190 parameter estimation: Application to a 1D ocean ecosystem model. *Journal of Marine Systems*,
 1191 89(1):1–18.
- 1192 Simon, E., Samuelson, A., Bertino, L., and Mouysset, S. (2015). Experiences in multiyear combined
 1193 state–parameter estimation with an ecosystem model of the North Atlantic and Arctic Oceans using
 1194 the ensemble Kalman filter. *Journal of Marine Systems*, 152:1–17.
- 1195 Skakala, J., Awty-Carroll, K., Menon, P. P., Wang, K., and Lessin, G. (2023). Future digital twins:
 1196 emulating a highly complex marine biogeochemical model with machine learning to predict hypoxia.
 1197 *Frontiers in Marine Science*, 10:1058837.
- 1198 Skakala, J., Bruggeman, J., Brewin, R. J., Ford, D. A., and Ciavatta, S. (2020). Improved represen-
 1199 tation of underwater light field and its impact on ecosystem dynamics: A study in the North Sea.
 1200 *Journal of Geophysical Research: Oceans*, 125(7):e2020JC016122.
- 1201 Skákala, J., Ford, D., Brewin, R. J., McEwan, R., Kay, S., Taylor, B., de Mora, L., and Ciavatta,
 1202 S. (2018). The assimilation of phytoplankton functional types for operational forecasting in the
 1203 Northwest European shelf. *Journal of Geophysical Research: Oceans*, 123(8):5230–5247.
- 1204 Skákala, J., Ford, D., Bruggeman, J., Hull, T., Kaiser, J., King, R. R., Loveday, B., Palmer, M. R.,
 1205 Smyth, T., Williams, C. A., et al. (2021). Towards a multi-platform assimilative system for North
 1206 Sea biogeochemistry. *Journal of Geophysical Research: Oceans*, 126(4):e2020JC016649.
- 1207 Skákala, J., Ford, D., Fowler, A., Lea, D., Martin, M. J., and Ciavatta, S. (2024). How uncertain and
 1208 observable are marine ecosystem indicators in shelf seas? *Progress in Oceanography*, 224:103249.
- 1209 Smith, V. H. and Schindler, D. W. (2009). Eutrophication science: where do we go from here? *Trends*
 1210 *in ecology & evolution*, 24(4):201–207.
- 1211 Smyth, T. J., Fishwick, J. R., Al-Moosawi, L., Cummings, D. G., Harris, C., Kitidis, V., Rees, A.,
 1212 Martinez-Vicente, V., and Woodward, E. M. (2010). A broad spatio-temporal view of the Western
 1213 English Channel observatory. *Journal of Plankton Research*, 32(5):585–601.
- 1214 Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J. (2016). Data assimilation in a coupled
 1215 physical–biogeochemical model of the California current system using an incremental lognormal 4-
 1216 dimensional variational approach: Part 1—model formulation and biological data assimilation twin
 1217 experiments. *Ocean Modelling*, 106:131–145.
- 1218 Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V. (2021). Bridging
 1219 observations, theory and numerical simulation of the ocean using machine learning. *Environmental*
 1220 *Research Letters*, 16(7):073008.
- 1221 Sonntag, S. and Hense, I. (2011). Phytoplankton behavior affects ocean mixed layer dynamics through
 1222 biological-physical feedback mechanisms. *Geophysical Research Letters*, 38(15).

- 1223 Telszewski, M., Palacz, A., and Fischer, A. (2018). Biogeochemical in situ observations—motivation,
1224 status, and new frontiers. *New Frontiers in Operational Oceanography*, pages 131–160.
- 1225 Teruzzi, A., Bolzon, G., Feudale, L., and Cossarini, G. (2021). Deep chlorophyll maximum and
1226 nutricline in the Mediterranean Sea: emerging properties from a multi-platform assimilated biogeo-
1227 chemical model experiment. *Biogeosciences*, 18(23):6147–6166.
- 1228 Teruzzi, A., Dobricic, S., Solidoro, C., and Cossarini, G. (2014). A 3-D variational assimilation scheme
1229 in coupled transport-biogeochemical models: Forecast of Mediterranean biogeochemical properties.
1230 *Journal of Geophysical Research: Oceans*, 119(1):200–217.
- 1231 Umlauf, L. and Burchard, H. (2011). Diapycnal transport and mixing efficiency in stratified boundary
1232 layers near sloping topography. *Journal of physical oceanography*, 41(2):329–345.
- 1233 Vagle, S., McNeil, C., and Steiner, N. (2010). Upper ocean bubble measurements from the NE Pacific
1234 and estimates of their role in air-sea gas transfer of the weakly soluble gases nitrogen and oxygen.
1235 *Journal of geophysical research: oceans*, 115(C12).
- 1236 van der Merwe, R., Leen, T. K., Lu, Z., Frolov, S., and Baptista, A. M. (2007). Fast neural network
1237 surrogates for very high dimensional physics-based models in computational oceanography. *Neural*
1238 *Networks*, 20(4):462–478.
- 1239 Wakelin, S. L., Artioli, Y., Butenschön, M., Allen, J. I., and Holt, J. T. (2015). Modelling the
1240 combined impacts of climate change and direct anthropogenic drivers on the ecosystem of the
1241 Northwest European continental shelf. *Journal of Marine Systems*, 152:51–63.
- 1242 Wakelin, S. L., Artioli, Y., Holt, J. T., Butenschön, M., and Blackford, J. (2020). Controls on near-
1243 bed oxygen concentration on the Northwest European Continental Shelf under a potential future
1244 climate scenario. *Progress in Oceanography*, 187:102400.