Learning-based calibration of ocean carbon models to tackle physical forcing uncertainties and observation sparsity

Jean Littaye, Ronan Fablet, Laurent Mémery

Learning-based calibration of ocean carbon models to tackle physical forcing uncertainties and observation sparsity.

J. Littaye[1,2], R. Fablet[1,3], L. Memery[2]

[1]IMT Atlantique, UMR Lab-STICC, Brest, France
[2]Laboratoire des sciences de l'Environnement MARin, UBO/CNRS/IRD/Ifremer, Plouzané, France
[3] ODYSSEY Team, INRIA, France

Key Points:

- Calibrating biogeochemical models through assimilation is sensitive to physical forcing uncertainties and to the observation configuration.
- The calibration of the biogeochemical models can be stated as a learning problem.
- The neural method shows a more robust calibration compared to data assimilation when dealing with imperfect forcings, sparse observations.

Corresponding author: J. Littaye, `Jean.Littaye@univ-brest.fr`

Abstract

Biogeochemical (BGC) ocean models are simplified representations of complex coupled processes, usually resulting in a large number of parameters, that need to be calibrated. In general, these parameters are constrained relying on incomplete and very heterogeneous sets of data. In addition, as biogeochemical tracers strongly depend on ocean circulation, the spatio-temporal uncertainties in the physical forcing can bias the circulation, which makes challenging the calibration of ocean carbon models. This study addresses the calibration of ocean biogeochemical models when dealing with imperfect physical forcings and sparse observations. We design a numerical testbed based on a simple 0D+t BGC model. It comprises different uncertainty scenarios for the physical forcing as well as different observation configurations of the considered NPZD (nutrient, phytoplankton, zooplankton, detritus) dynamics. We propose and benchmark a learning-based scheme against a variational data assimilation (DA) approach. The former frames the calibration as learning a neural operator between observations and model parameters. The experiments revealed that the DA-based calibration is highly sensitive to imperfect physical forcing and limited observations, often leading to significant estimation errors in BGC parameters. Conversely, the learning-based approach demonstrated a greater robustness in parameter estimation and simulated BGC patterns. We discuss further how these results could transfer to more realistic BGC models and real observing systems.

Plain Language Summary

Ocean carbon models are essential tools for studying climate change, especially its role in the earth's carbon cycle. However, calibrating these models requires diverse data sources, including observational datasets that are often scarce and model-based datasets that contain significant uncertainties. The quality of these data impacts the accuracy and validity of the models. This study explores the relationship between data quality and model validity using a simple ocean carbon model. Additionally, the study investigates emerging learning methods, as neural networks, that handle imperfect data to represent ocean processes. By comparing a traditional variational data assimilation method with a new learning-based approach, the study evaluates their effectiveness in model calibration. The results show that the traditional method is highly sensitive to data quality, while the learning method is more robust. As these results are representative of an idealised framework with a simple carbon model, we conclude by discussing how this method could apply to more realistic models.

Index terms and keywords

Index terms:

1. **0414** Biogeochemical cycles, processes, and modeling.
2. **0555** Neural networks, fuzzy logic, machine learning.
3. **1990** Uncertainty.
4. **4260** Ocean data assimilation and reanalysis.
5. **4273** Physical and biogeochemical interactions.

Keywords: Calibration, Physical forcing uncertainty, Sparse observation, Neural mapping, Data assimilation, Biogeochemical model

1 Introduction

The ocean is an extremely rich and diversified ecosystem, that also represents the largest carbon reservoir evaluated to $40000 PgC$ (petagram $= 10^{15} g$ of carbon) stored in the deep ocean (Williams & Follows, 2011). It also plays a key role in the carbon cy-

cle, by exchanging carbon dioxide (CO_2) with the atmosphere. Yet, the monitoring and estimation of the carbon fluxes at the air-sea interface as well as between the upper ocean and the deep ocean remains highly challenging (Henson et al., 2022). They are among the major sources of uncertainties for the characterization of the overall carbon cycle, especially under climate change and the resulting global warming of the ocean (Sarmiento et al., 2004).

Besides dedicated observing systems such as the Global Ocean Observing System (https://goosocean.org/), the design of accurate ocean biogeochemical (BGC) models is a key component of earth system models to estimate carbon fluxes in the ocean and between the ocean and the atmosphere. The advances in the parameterization of complex physical and biogeochemical interactions have resulted in an increasing complexity of ocean BGC models (Ismail & Al-Shehhi, 2022; Friedrichs et al., 2007). In turn, the calibration of these models has become more challenging. Besides its computational cost, the complexity of the calibration problem involves two main aspects: the scarcity of the available ocean observation datasets, the uncertainties and biases in the physical forcing. Regarding ocean BGC variables, even for sea surface dynamics, the sensitivity of satellite sensors to the cloud cover result in large missing data rates on a global scale (IOCCG, 2007). In situ networks, including ARGO floats and moored buoys, provide observation dataset for the ocean's interior with daily-to-monthly characteristic time samplings, but their spatial coverage is very scarce (Gould et al., 2013; Claustre et al., 2020). Ocean circulation is one of the main physical variables driving ocean BGC processes (Fossette et al., 2012; Berline et al., 2007; Claustre et al., 2021). Therefore, sampling gaps for ocean dynamics also result in significant uncertainties in physical ocean reanalyses, especially for mesoscale and submesoscale dynamics which play a key role in ocean BGC dynamics (Picard, Gula, et al., 2024; Lévy, 2008; McGillicuddy Jr et al., 1998). For instance, as stressed in Ballarotta et al. (2019), operational products cannot inform physical dynamics at sea surface for horizontal scales below ≈ 100 km. From a methodological point of view, the calibration of ocean BGC models generally leverages a data assimilation formulation to solve jointly for the reconstruction of the ocean BGC processes as well as for the estimation of the parameters of the ocean BGC model (Dowd et al., 2014). By considering solely the ocean BGC component, the calibration method likely suffers from the above-mentioned uncertainties in the physical forcing, associated with available ocean reanalyses (Pasquier et al., 2023; Doney et al., 2004). While considering the assimilation of forcing, physical and biogeochemical parameters in a coupled ocean physical-biogeochemical model may seem appealing to address this shortcoming, it remains computationally prohibitive and numerically challenging (Kane et al., 2011) for operational systems.

Learning-based approaches and especially Deep Learning (DL) approaches have recently gained interest in ocean science for a variety of topics including among others the processing of in situ data, the reconstruction of gap-free fields from sparse observation datasets (Fablet, Amar, et al., 2021), the calibration of closure terms (Gupta & Lermusiaux, 2021), the short-term forecasting of ocean states (Yin et al., 2021; Fablet, Amar, et al., 2021). The generic feature of these approaches is to state the considered problem as the learning of an operator to map some input data, typically observation data, to the targeted variable. Some studies involve applications to real observation datasets (Roussillon et al., 2023; Febvre et al., 2023; S. A. Martin et al., 2024) which emphasize the readiness of state-of-the-art DL approaches to scale up to the complexity of real ocean datasets and processes (Bolton & Zanna, 2019), including regarding observation noise, modeling errors and sampling gaps.

These recent advances support the potential of DL schemes to address calibration challenges for ocean BGC models. Here, we explore this learning methodology and state the targeted calibration problem as the training of a neural mapping between observation data and model parameters. We design an experimental framework for evaluation purposes and benchmark the proposed approach with respect to a data-assimilation-based

calibration scheme. In particular, we used a variational data assimilation method as a reference calibration method. We rely on an idealized 0D+t case-study accounting for different scenarios for physical forcing uncertainties and observation configurations. Our objective is twofold: (i) to evaluate the impact of physical forcing uncertainties on a DA-based calibration method of a BGC model, (ii) to explore whether a learning-based calibration method can improve BGC parameter estimation when accounting for these uncertainties. We discuss further the significance of these results for the real ocean, especially regarding the deployment of neural calibration schemes to observation datasets.

This paper is organized as follows. Section 2 introduces the targeted problem and the considered methodology. Section 3 describes the BGC model and the associated evaluation framework for calibration purposes. We report our numerical experiments in Section 4. Section 5 discusses our results and future work.

2 Methods

This section first states the calibration of ocean BGC models as an inverse problem given partial and noisy observations and error-prone physical forcings. We then introduce a data-assimilation-based method as well as the proposed learning-based scheme. We also detail the considered evaluation framework to assess the performance of the calibration methods under different observation and uncertainty scenarios.

2.1 Problem statement

This study addresses the estimation of the parameters of an ocean BGC model given sparsely-sampled and noisy observations of BGC variables and physical forcings that are subject to uncertainties. Formally, let us introduce $\mathbf{X}(t)$ the ocean BGC state (with $\mathbf{X_0} = \mathbf{X}(0)$ the initial state), $\mathbf{U}(t)$ the physical forcings and $\mathbf{Y}(t)$ the observed variable at time t. The considered calibration problem relates to the following state-space formulation:

$$\begin{cases} \dfrac{d\mathbf{X}(t)}{dt} = \mathcal{M}_\theta(\mathbf{X}(t), \mathbf{U}(t)) \\ \mathbf{Y}(t) = \mathcal{H}(\mathbf{X}(t)) + \nu_\mathbf{s}(t) \end{cases} \tag{1}$$

where $\mathcal{M}_\theta$ is the ocean BGC model. It describes the time evolution of the ocean BGC state, typically according to an ordinary or partial differential equation (Aumont et al., 2015; Ismail & Al-Shehhi, 2022; Fennel et al., 2022). $\mathcal{H}(.)$ is the observation operator. It accounts for the space-time sampling of the considered observing systems. Importantly, for ocean BGC processes, in situ and satellite-derived observations result in a very scarce sampling. $\nu_\mathbf{s}$ refers to the observation noise, in general stated as a white Gaussian noise. In (1), variable $\mathbf{U}$ denotes the physical forcings, which drive the BGC dynamics, such as light, temperature or mixing (non-exhaustive list). (Dowd et al., 2014).

Operational reanalysis products (European Union-Copernicus Marine Service, 2019) leverage data assimilation schemes and available observation datasets and deliver estimates of the forcings, denoted as $\mathbf{U}^\star$. Given the computational complexity of the associated inverse problem and the scarcity of the observations, especially for the ocean component, these estimations may involve significant uncertainties and biases (Lin et al., 2022; Park et al., 2018; Lermusiaux et al., 2006; Doney et al., 2004). Let us denote by $\mathbf{U}^\star = \mathbf{U} + \nu_\mathbf{U}$ with $\nu_\mathbf{U}$ a time-dependent error on the physical forcings.

This study aims at retrieving the parameters of the BGC model θ given observation data $\mathbf{Y}$ and error-prone forcings $\mathbf{U}^\star$. Whereas data assimilation schemes (Dowd et al., 2014; Kane et al., 2011; Berline et al., 2007) provide a general methodological framework to address this issue, learning-based paradigms also appear as appealing strategies to solve inverse problems in geoscience (Frerix et al., 2021; Zhang et al., 2020; Farchi et

al., 2023; Bonavita & Laloyaux, 2020). We explore these two frameworks in the following sections.

2.2 DA-based calibration scheme

Among the state-of-the-art data assimilation (DA) methods (Pasquier et al., 2023; Park et al., 2018; Kane et al., 2011; Brasseur et al., 2009; Berline et al., 2007), this study focuses on a variational data assimilation scheme to the considered BGC model calibration problem. To account for model error, we consider a weakly-constrained 4DVar scheme (Trémolet, 2007; Fablet, Chapron, et al., 2021; Frerix et al., 2021). It comes to solve the following minimization problem w.r.t. model parameters θ and BGC state $\mathbf{X}$ given some observation data $\mathbf{Y}$ and the estimated physical forcings $\mathbf{U}^\star$:

$$\begin{aligned}\mathcal{J}(\mathbf{X},\mathbf{Y},\theta,\mathbf{U}^\star) = \sum_k \left(\sum_i \|\mathcal{H}(\mathcal{T}_{\mathcal{M}_\theta,\delta}(\mathbf{X}(k\Delta+(i-1)\delta),\mathbf{U}^\star)) - \mathbf{Y}(k\Delta+i\delta)\|_R^2 \right) \\ +\Lambda \sum_k \|\mathcal{T}_{\mathcal{M}_\theta,\Delta}(\mathbf{X}(k\Delta),\mathbf{U}^\star) - \mathbf{X}((k+1)\Delta)\|_B^2\end{aligned} \tag{2}$$

where $\mathcal{T}_{\mathcal{M}_\theta,\Delta}$ is the time-stepping operator of the BGC model $\mathcal{M}_\theta$ in (1) such that $\mathcal{T}_{\mathcal{M}_\theta,\Delta}(\mathbf{X_0},\mathbf{U}^\star)$ is the Δ-step-ahead prediction of the BGC state from initial condition $\mathbf{X_0}$ and physical forcings $\mathbf{U}^\star$. The integer $k \in [\![1,K]\!]$ refers to the number of Δ-size time periods, that can be divided into I sub-time periods of length δ, s.t. $\Delta = I\delta$ and $i \in [\![1,I]\!]$. For a studied period T homogeneously time-sampled with N_T samples, we have $T = K\Delta = KI\delta = N_T\delta$. In the cost function (Equation 2), the first term assesses the consistency between the reconstructed BGC state $\mathbf{X}$ and the observation $\mathbf{Y}$ while the second term evaluates the consistency of the BGC state w.r.t. dynamical model $\mathcal{M}_\theta$ over a given time window Δ. R and B refer to the observation covariance and the model error covariance respectively, while the product $||u||_A^2 = u^T A u$ with u a vector and A a matrix. The observation covariance matrix $R = \sigma_{obs}^2 \mathbb{I}$ with $\mathbb{I}$ the identity matrix and where σ_{obs} is the standard deviation of observation noise $\nu_{\mathbf{s}}$. The model error covariance matrix $B = \sigma_{mod}^2 \mathbb{I}$ with σ_{mod} the standard deviation of observations along the T period. An additional term Λ weighs the relative contribution of the dynamical prior. We also include prior knowledge on the parameter space assuming a uniform distribution. It is implemented as a hard constraint in the considered minimization and prevents the estimated model parameters from taking unrealistic values.

As generally assumed in operational systems (Guiavarc'h et al., 2019; Park et al., 2018), we consider in minimization (Eq. 2) the estimated forcings $\mathbf{U}^\star$ in place of the true forcing $\mathbf{U}$. As such, this DA-based formulation only accounts for the uncertainty on the physical forcings through the model error covariance. To better address these uncertainties, one could state a DA issue for a coupled physics-BGC system as explored in (Pasquier et al., 2023; Park et al., 2018; Berline et al., 2007; Fennel et al., 2001). This greatly increases the computational complexity of the associated inverse problem, which impedes its exploitation for real case-studies (Dowd et al., 2014).

Numerically speaking, we assume that we are provided with the adjoint operator or a differentiable implementation of model $\mathcal{M}$ with respect to parameters and states. Then we solve the above minimization with a fixed-step gradient descent. We initialize this gradient descent randomly for parameters θ, using observed states when they are available and an average value of state $\overline{\mathbf{X}}$ (averaged from all the available realizations) for the unobserved components of $\mathbf{X}$. We tune weighing parameter Λ empirically through a grid search procedure for the considered experiments. The step of the gradient descent is set to ensure the convergence of the minimization and we typically run 10000 gradient steps.

Among the variational methods, ensemble variational methods are particularly advantageous in that they enable to consider various members of a distribution simultaneously. They are typically used to estimate the uncertainty of estimated variables (Bannister, 2017). In this study, an ensemble of L realizations of estimated forcing $\mathbf{U}^{(j,l)\star} = \mathbf{U}^{(j)} + \nu_{\mathbf{U}}{}^{(j,l)}$ is considered for each sample $j \in [\![1, J]\!]$, with $l \in [\![1, L]\!]$ the number of the member. Assuming the forcing uncertainty distribution is centred on 0 (Bannister, 2017), i.e. $\frac{1}{L}\sum_l \nu_{\mathbf{U}}{}^{(l)} = 0$, the estimated forcing $\mathbf{U}^\star$ is non-biased, and $\frac{1}{L}\sum_l \mathbf{U}^{(j,l)\star} = \frac{1}{L}\sum_l (\mathbf{U}^{(j)} + \nu_{\mathbf{U}}{}^{(j,l)}) = \mathbf{U}^{(j)}, \forall j \in [\![1, J]\!]$. Therefore, if the calibration method is a linear process denoted by $\mathcal{K}(.)$ such that $\theta = \mathcal{K}(\mathbf{Y}, \mathbf{U}^\star)$, then the BGC parameters can be obtained independently of the forcing draw, assuming a sufficiently large and representative ensemble. According to the previous equation, the actual BGC parameter of a sample j is the average of the L estimated parameters.

2.3 Learning-based calibration scheme

Besides the DA-based method, we propose a supervised learning calibration scheme. It leverages Observing Simulation System Experiments (OSSEs) to build representative training and validation datasets (Febvre et al., 2023; Fablet, Amar, et al., 2021). It states the calibration problem as the learning of a neural mapping $\mathcal{K}_\Theta$ from observations $\mathbf{Y}$ and estimated forcings $\mathbf{U}^\star$ to BGC model parameters θ (Figure 1). Θ refers to the parameters of the neural network. Ideally, an error-free mapping would lead to $\theta = \mathcal{K}_\Theta(\mathbf{Y}, \mathbf{U}^\star)$. We benefit from the versatility of neural networks to define mapping operators between tensor spaces of different dimensions.

We explore three neural architectures with varying complexity: a fully-connected neural network, referred to as multi-layer perceptron (MLP), a convolutional neural network (CNN) and a U-Net. The MLP and CNN are simple baseline neural architectures (LeCun et al., 2015). Numerous studies acknowledge better performance of CNN architectures, including for ocean applications (Smith et al., 2023; Roussillon et al., 2023; Ducournau & Fablet, 2016), when dealing with nD-dimensional tensors such as multi-variate time series, images and space-time fields. U-Net architectures (Ronneberger et al., 2015) can be regarded as a multi-scale extension of CNNs to better account for multi-scale patterns in multi-dimensional tensors. They are among the state-of-the-art architectures or elementary components for many applications and methods including among others signal and image segmentation, image-to-image mapping, diffusion models. (Luo et al., 2023). We can also cite various applications in ocean studies (Du & Zhang, 2024; Picard, Baker, et al., 2024). We illustrate the computational graph of these architectures in Figure 1, that have been designed for the study. Through these three architectures, we aim to assess whether the performance of learning-based schemes strongly depends on the considered neural method, independently of any architecture optimization process. All these architectures share the same inputs and outputs. The concatenation of observations $\mathbf{Y}$ and forcings $\mathbf{U}^\star$ form the $N_{ch} \times N_T$ input tensor, for $N_{ch} = N_{state} + N_{forcing}$ with N_{state} the number of observed states, $N_{forcing}$ the number of physical forcings and N_T the number of observed time-steps, while the output tensor is a H-dimensional vector, where the BGC model contains H BGC parameters. We apply a linear interpolation to the observed states $\mathbf{Y}$, so that the resulting observation tensor matches the time dimension of the forcing tensor.

For the training phase, we assume that we are provided with a representative set of OSSEs. Let us denote by $\{\mathbf{Y}^{(j)}, \mathbf{X}^{(j)}, \mathbf{U}^{(j)}, \mathbf{U}^{\star(j)}, \theta^{(j)}\}_{j=1,\dots,N}$, where index j refers to each individual OSSE. Given this dataset, we learn the parameters of the neural mapping $\mathcal{K}_\Theta$ from the minimization of the following supervised training loss:

$$\arg\min_\Theta \sum_j \|\theta^{(j)} - \mathcal{K}_\Theta(\mathbf{Y}^{(j)}, \mathbf{U}^{\star(j)})\|_\Sigma^2 \tag{3}$$

where covariance Σ encodes the relative weight given to each BGC parameter in θ. We consider a diagonal covariance to normalize the parameter calibration error according to the variance of each parameter in the training dataset. The training phase exploits a stochastic gradient descent as classically used in deep learning approaches (Bottou, 2010). We use Adam optimizer with a learning rate of $1e-4$ and random mini-batches with 512 samples.

2.4 Evaluation framework

For evaluation purposes, we assess the performance of the considered calibration schemes according to two main categories of metrics: metrics defined in the model parameter space and metric defined in the BGC state space. Hereafter, we denote by $\widehat{\theta} \in \{\theta^{DA},\ \theta^{NN}\}$ the estimated BGC model parameters θ using either the DA-based (θ^{DA}) or the NN-based (θ^{NN}) calibration scheme and $\overline{\theta}$ the associated average over the J samples.

The first category of metrics comprises the normalized root mean square error (NRMSE) and normalized mean error (NME) for each of the h parameters:

$$\begin{aligned} NRMSE(\widehat{\theta}_h) &= \frac{1}{J}\frac{\sum_j \left|\widehat{\theta}_h^{(j)} - \theta_h^{(j)}\right|}{\overline{\theta}_h}, h \in [\![1, H]\!] \\ NME(\widehat{\theta}_h) &= \frac{1}{J}\frac{\sum_j \widehat{\theta}_h^{(j)} - \theta_h^{(j)}}{\overline{\theta}_h}, h \in [\![1, H]\!] \end{aligned} \tag{4}$$

as well as over all parameters:

$$\begin{aligned} NRMSE(\widehat{\theta}) &= \frac{1}{H \times J}\sum_h \sum_j \frac{\left|\widehat{\theta}_h^{(j)} - \theta_h^{(j)}\right|}{\overline{\theta}_h} \\ NME(\widehat{\theta}) &= \frac{1}{H \times J}\sum_h \sum_j \frac{\widehat{\theta}_h^{(j)} - \theta_h^{(j)}}{\overline{\theta}_h} \end{aligned} \tag{5}$$

The subscripts j and h denote the number of the sample considered in the dataset and the number of the BGC parameter of the model, respectively, i.e. $\theta_h^{(j)}$ is the h-th parameter of the j-th sample. We denote by θ^{DA} estimated parameters for the DA-based method (θ^{NN} for the NN-based method).

Regarding the metrics in the BGC state space, we proceed as follows. For sample j in the test dataset, we compare the BGC state time series simulated using initial condition $\mathbf{X_0}$ and estimated physical forcings $\mathbf{U}^\star$ using both estimated parameters $\widehat{\theta}$ and true ones θ. The notation $\mathbf{X}$ refers to BGC states simulated with actual parameters θ, while $\widehat{\mathbf{X}} \in \{\mathbf{X}^{DA},\ \mathbf{X}^{NN}\}$ denotes the states simulated with estimated parameters $\widehat{\theta} \in \{\theta^{DA},\ \theta^{NN}\}$. We compute three other metrics that assess the patterns, variations, errors as Séférian et al. (2013) as well as the shifts:

- **The correlation** $Corr(\mathbf{X}, \widehat{\mathbf{X}})$ that is either applied between $\mathbf{X}$ and $\mathbf{X}^{DA}$ or between $\mathbf{X}$ and $\mathbf{X}^{NN}$.
- **The shift** $|\arg\max_{dt}\{corr(\widehat{\mathbf{X}}(t), \mathbf{X}(t+dt))\}|$, defined (in days) as the shift between two signals where the correlation is the highest. It is either calculated between $\mathbf{X}$ and $\mathbf{X}^{DA}$ or between $\mathbf{X}$ and $\mathbf{X}^{NN}$.

- **The amplitude ratio** $AR = 2 \times \frac{\max(\widehat{\mathbf{X}}) - \min(\widehat{\mathbf{X}})}{\max(\widehat{\mathbf{X}}) - \min(\widehat{\mathbf{X}}) + \max(\mathbf{X}) - \min(\mathbf{X})}$ that features the ratio between the amplitude (defined as the difference between the maximum and the minimum) of two signals during the studied period. It is either calculated between $\mathbf{X}$ and $\mathbf{X}^{DA}$ or between $\mathbf{X}$ and $\mathbf{X}^{NN}$.

The closer a signal to the reference, the closer to 1.0 (resp. 0.0 and 1.0) Corr (resp. Shift and AR). Among the three metrics, the correlation represents the similarity between the state time series using estimated and true BGC model parameters. The shift indicates potential temporal biases. The ratio AR reveals differences in the amplitude of $\widehat{\mathbf{X}}$ compared to $\mathbf{X}$: values close to 0 indicates an underestimation of $\mathbf{X}$ and values close to 2 an overestimation. We illustrate in Fig.2 these four metrics to compare two zooplankton concentration times series.

We compute the above criteria for all the samples of an evaluation dataset. Overall, we evaluate their mean values and analyse their distributions through violin plots, introduced by Hintze and Nelson (1998).

3 Case-study and Data

This section details the considered 0D+t ocean BGC case-study. It comprises the description of the chosen BGC model, the specification of physical forcing uncertainties and observation configurations, as well as a synthesis of the resulting datasets used both for training and evaluation purposes.

3.1 The BGC model

We consider a 0D+t BGC model for the time evolution of the nitrogen concentration through 4 compartments, namely nutrients ($\mathbf{N}$), phytoplankton ($\mathbf{P}$), zooplankton ($\mathbf{Z}$) and detritus ($\mathbf{D}$). From Brady (2017), we added mixing and sedimentation of detritus. The model then relies on the following coupled ordinary differential equations:

$$\begin{cases} \frac{d\mathbf{N}}{dt} = -\psi(\frac{\mathbf{N}}{\chi+\mathbf{N}})\mathbf{fP} + \alpha\rho(1-e^{-\lambda\mathbf{P}})\mathbf{Z} + \epsilon\mathbf{P} + \gamma\mathbf{Z} + \varphi\mathbf{D} + \mathbf{m}(Q_s - \mathbf{N}) \\ \frac{d\mathbf{P}}{dt} = \psi(\frac{\mathbf{N}}{\chi+\mathbf{N}})\mathbf{fP} - \rho(1-e^{-\lambda\mathbf{P}})\mathbf{Z} - \epsilon\mathbf{P} - \eta\mathbf{P} - \mathbf{mP} \\ \frac{d\mathbf{Z}}{dt} = \beta\rho(1-e^{-\lambda\mathbf{P}})\mathbf{Z} - \gamma\mathbf{Z} - \mathbf{mZ} \\ \frac{d\mathbf{D}}{dt} = \eta\mathbf{P} + (1-\alpha-\beta)\rho(1-e^{-\lambda\mathbf{P}})\mathbf{Z} - \varphi\mathbf{D} - \zeta\mathbf{D} - \mathbf{mD} \end{cases} \tag{6}$$

For this system, we define the BGC state as $\mathbf{X} = \begin{pmatrix}\mathbf{N} & \mathbf{P} & \mathbf{Z} & \mathbf{D}\end{pmatrix}$ as the state variables, the physical forcings as $\mathbf{U} = (\mathbf{f}\ \mathbf{m})$, and $\theta = (\psi\ \chi\ \ldots\ \zeta)$ comprises all model parameters. We sum up this information in Tab.1. According to notations introduced in Section 2.3, $N_{ch} = 4 + 2$ and $H = 11$. The considered parameterization combines the default parametrisation described in https://github.com/bradyrx/NPZD-Model/blob/master/NPZD-Model.ipynb and empirical cross-validation experiments to simulate realistic patterns, especially accounting for a main phytoplankton bloom during spring followed by lower detritus and zooplankton peaks, with few oscillations.

This BGC model gathers several sub-processes appearing in one or several equations of Eq.6:

- $\psi(\frac{\mathbf{N}}{\chi+\mathbf{N}})\mathbf{fP}$ the nutrient uptake of phytoplankton that relies on the quantity of phytoplankton, the light and nutrients availability.

Symbol	Definition	Reference value
ψ	Maximum growth rate of the phytoplankton	1 d^{-1}
χ	Half-saturation constant for nitrogen uptake	1 $\mu molN.L^{-1}$
ϵ	Phytoplankton respiration rate	0.1 d^{-1}
η	Phytoplankton death rate	0.15 d^{-1}
ρ	Maximum zooplankton grazing rate on phytoplankton	2 d^{-1}
λ	Zooplankton grazing constant	0.05 $\mu molN.L^{-1}$
β	Proportion of assimilated nitrogen by zooplankton	0.6 (*dimensionless*)
α	Proportion of nitrogen taken up by zooplankton that returns to the environment as dissolved nutrients	0.3 (*dimensionless*)
γ	Zooplankton death rate	0.1 d^{-1}
φ	Detritus remineralization rate	0.4 d^{-1}
ζ	Sedimentation rate	0.1 d^{-1}
$\mathbf{f}$	Light limitation (from 0 to 1)	(*dimensionless*)
$\mathbf{m}$	Vertical mixing at the base of the 0D+t box	d^{-1}

Table 1. Table of the BGC parameters (the first 11), the physical forcings (the last 2), their definition and reference value.

- $\rho(1-e^{-\lambda \mathbf{P}})\mathbf{Z}$ the zooplankton grazing that relies on quantity of zooplankton, their grazing ability and the quantity of phytoplankton.
- $\epsilon\mathbf{P}$ the respiration of the phytoplankton.
- $\gamma\mathbf{Z}$ the release of nitrogen from dead zooplankton.
- $\varphi\mathbf{D}$ the remineralization of the detritus.
- $\mathbf{m}(Q_s-\mathbf{N})$ the nutrient supply by mixing from a source $Q_s = 8\mu molN.L^{-1}$ supposed constant in a lower layer.
- $\mathbf{mP/Z/D}$ the export of phytoplankton, zooplankton and detritus implied by the mixing toward the lower layer.
- $\eta\mathbf{P}$ the release of nitrogen from dead phytoplankton.
- $\zeta\mathbf{D}$ the sedimentation of the sinking detritus.

For the two first terms (relative to the nutrient uptake and the zooplankton grazing), a lower threshold has been set for the phytoplankton and zooplankton (with minimum concentrations equal to $10^{-2}\mu mol.L^{-1}$). This threshold avoids a regime change (Scheffer et al., 2001) in which all the nitrogen remains in **N** while **P**, **Z** and **D** are zero, which is not relevant (weak interaction between states and forcings). We report the reference values of the model parameters in Tab.1.

The considered physical forcing account for light and mixing conditions. The light intensity **f** is a dimensionless value, where a value of 0 features no availability of light whereas 1 means no light limitation. The mixing **m** represents different mixing processes that acts in the nutrient supply such as internal waves, turbulence, vertical mixing (Williams & Follows, 2011; Circulation, 1989; Pickard & Emery, 1961).

3.2 Physical forcings

The physical forcings in Eq.7 involve the light limitation **f** and the mixing **m** that are time-dependent. We model these forcings as the sum of a seasonal pattern $\mathbf{a_f}$ for the light (resp. $\mathbf{a_m}$ for the mixing) and of a high-frequency (HF) component $\mathbf{b_f}$ (resp. $\mathbf{b_m}$).We consider a combination of seasonal sine patterns and of first-order auto-regressive processes with a hourly time resolution as follows:

$$
\begin{cases}
\mathbf{f}(t_i) = \mathbf{a_f}(t_i) + \mathbf{b_f}(t_i) \\
\mathbf{a_f}(t_i) = \dfrac{(f_{max} - f_{min})}{2} cos(2\pi \dfrac{t_i}{T_f} + \delta_f) + \dfrac{(f_{max} + f_{min})}{2} \\
\mathbf{b_f}(t_{i+1}) = \alpha_f \mathbf{b_f}(t_i) + \mathbf{c_f}(t_i),\ \mathbf{c_f} \sim N(0, \sigma_f^2) \\
\mathbf{m}(t_i) = \mathbf{a_m}(t_i) + \mathbf{b_m}(t_i) \\
\mathbf{a_m}(t_i) = \dfrac{(m_{max} - m_{min})}{2} cos(2\pi \dfrac{t_i}{T_m} + \delta_{m/f} + \pi + \delta_f) + \dfrac{(m_{max} + m_{min})}{2} \\
\mathbf{b_m}(t_{i+1}) = \alpha_m \mathbf{b_m}(t_i) + \mathbf{c_m}(t_i),\ \mathbf{c_m} \sim N(0, \sigma_\Delta^2) \\
t_{i+1} = t_i + dt, \forall i \in \mathbb{N}, dt \in \mathbb{R}
\end{cases}
\tag{7}
$$

The seasonal patterns are given by sine waves characterized by a maximum (resp. minimum) of amplitude f_{max} and m_{max} (resp. f_{min} and m_{min}), a fixed period $T_f = T_m$ and a shift δ_f and δ_m. In our experiments, the shift of the light δ_f is a fixed constant and leads to a maximum (resp. minimum) light limitation during summer (resp. winter) period. The mixing involves roughly an opposite seasonal pattern. We define the minimum seasonal values for the light limitation, denoted as f_{min}, as well as for mixing, denoted as m_{min}. The two physical forcings compete, i.e. when the light limitation reaches its maximum value, the mixing tends to be at its lowest value (with some delay $\delta_{m/f}$). This is relative to the fact that an increase of light goes with an increase of temperature that leads to a more stratified upper layer, then a lower nutrient supply. Due to different processes especially the thermal inertia of the ocean (Royce & La, 2011), there is a time lag $\delta_{m/f}$ between these two parameters. In our experiments, we vary the value of parameters f_{max}, m_{max} and $\delta_{m/f} = \delta_m - \delta_f - \pi$ around reference values ($\pm 20\%$) according to uniform random distributions. The variation of those values characterizes different environments of the ocean.

The HF components of the forcings account for smaller time-scale processes, typically with characteristic scales from a few days to a few weeks. Among others, the related uncertainties relate to the dynamics of the cloud coverage for the light and the (sub)mesoscale dynamics for the mixing (mainly eddies dynamics, Lévy (2008)). We consider first-order linear auto-regressive processes $\mathbf{b_{i \in \{f,m\}}}$, with α_i the slope and $\mathbf{c_i} \sim N(0, \sigma_i^2)$ a centred Gaussian white noise with variance σ_i^2. We simulate the true and estimated forcing, $\mathbf{U}$ and $\mathbf{U}^\star$, as two realizations of model (7) for the same parameter settings.

3.3 Simulated datasets

We perform numerical experiments with BGC model (6) and forcings (7) for different physical forcing uncertainties and observation configurations. We consider three uncertainty levels and vary the parameters of the high-frequency random component, namely α_f, α_m, σ_f and σ_m as follows:

- **Case-1:** The light and mixing HF variations have a standard deviation up to 4% of their maximum seasonal amplitudes. These HF components depict a period of about 7 days for $\mathbf{f}$ and about 3 days for $\mathbf{m}$ according to $\alpha_f = 0.9$, $\alpha_m = 0.8$, $\sigma_f = 0.001$ and $\sigma_m = 1.6e-5$.
- **Case-2:** The light and mixing HF variations have a higher standard deviation than Case-1. We set $\alpha_f = 0.9$, $\alpha_m = 0.8$, $\sigma_f = 0.0025$ and $\sigma_m = 4e-5$. This leads to an amplitude up to 6% of their maximum seasonal amplitudes. The period of the HF component is about 7 days for $\mathbf{f}$ and about 3 days for $\mathbf{m}$.
- **Case-3:** The light and mixing HF variations have the same standard deviation as Case-2 but a lower frequency. We set $\alpha_f = 0.95$, $\alpha_m = 0.9$, $\sigma_f = 0.001$ and $\sigma_m = 1.6e-5$. Those values characterize amplitudes up to 4% of their maxi-

Symbol	Value(s)
f_{max}	0.7 ± 0.14
m_{max}	0.1 ± 0.02 (day^{-1})
$\delta_{m/f}$	$\frac{\pi}{5} \pm \frac{\pi}{25}$
α_f	$\{0.9, 0.95\}$
σ_f	$\{0.001, 0.0025\}$
α_m	$\{0.8, 0.9\}$
σ_m	$\{1.6e-5, 4e-5\}$
f_{min}	0.05
m_{min}	0.02 (day^{-1})
$T_f = T_m$	365 (day)
δ_f	$2\pi\frac{20}{365}$

Table 2. Table of the physical forcing parameters and their reference value. The value of f_{max}, m_{max} and $\delta_{m/f}$ is constant for each of the generated OSSEs (see Section 3.3) and selected within the uniform distribution. A value of $a \pm b$ means that the variable is uniformly distributed between $a - b$ and $a + b$. Variables α_f, α_m, σ_f and σ_m are defined according to the considered case of uncertainty, defined Section 3.3.

mum seasonal amplitudes (for both **f** and **m**). Nonetheless, these signals have a period of about 10 days for **f** and about 7 days for **m**.

We recall that the true and estimated forcings $\mathbf{U}$ and $\mathbf{U}^\star$ involve two realizations of (7) with the same parameterization. An additional **Case-0** is also considered in one of our experiments with the statistics of Case-1 but where $\mathbf{U} = \mathbf{U}^\star$, which means that we perfectly know the forcings. We report in Fig.3 examples of the simulated physical forcing and of the associated BGC simulations for a Case-3 parameterization.

Regarding the observation configurations, we draw inspirations from real observing systems to define three observation schemes. In general, nutrients and phytoplankton are more easily and frequently sampled than zooplankton and detritus. In our experiments, the three observation schemes differ in the sampling rates considered for the four BGC variables as follows:

- Scheme 1/7: we assume a daily sampling for the concentrations of nutrients and phytoplankton and a weekly sampling for zooplankton and detritus.
- Scheme 7/7: we assume a weekly sampling for the concentrations of nutrients, phytoplankton, zooplankton and detritus.
- Scheme 1/0: we assume a daily sampling for the concentrations of nutrients and phytoplankton, while we do not directly observe zooplankton and detritus.

Our observation configurations also account for an additive Gaussian noise for each BGC variable with a standard deviation set to 1% of the annual variability standard deviation, i.e. ± 0.030 (resp. 0.015, 0.005 and 0.004) $\mu mol.L^{-1}$ for **N** (resp. **P**, **Z** and **D**).

We generate different datasets for each uncertainty level. Each dataset involves $N + J = 5100$ simulations of the BGC model, each simulation being run over five years with parameters selected uniformly within a 20% interval around the reference value given in Tab.1 and a true physical forcing simulated from Eq.7 according to the considered uncertainty level. We consider a split between training, validation and test datasets to en-

sure their statistical independence as classically used in deep learning studies (Xu & Goodacre, 2018). Among the 5100 simulations, $N = 5000$ simulations are kept to train the NN, i.e. 4000 simulations are kept for the training dataset, 1000 for the validation dataset, and $J = 100$ for the test dataset (see Fig.4). All simulations share the same initial condition given by:
$\mathbf{X_0} = \big(\mathbf{N_0} = 4 \quad \mathbf{P_0} = 2.5 \quad \mathbf{Z_0} = 1.5 \quad \mathbf{D_0} = 0\big)\, \mu mol.L^{-1}$. As a trade-off between accuracy and calculation time, we use the 1st-order explicit Euler method with an integration step of 12 hours. For each simulation, we extract the third year as a reference calibration sample with true and error-prone forcing time series $\mathbf{U}$ and $\mathbf{U}^\star$, true BGC state time series $\mathbf{X}$ and true model parameters θ. Besides for the test dataset, we use the fifth year to compute BGC state metrics as defined in Section 2.4.

Concerning the ensemble method, several realizations of the forcing, denoted $\{\mathbf{U}^{(l)\star}\}_{l\in L}$ were generated. For each of the 100 simulations of the test dataset, $L = 100$ realizations of the forcing are considered. For each realization $\mathbf{U}^{(j,l)\star} = \mathbf{U}^{(j)} + \nu_{\mathbf{U}}{}^{(j,l)}$, the seasonal signal $\mathbf{U}^{(j)}$ remains the same but the uncertainty $\nu_{\mathbf{U}}{}^{(j,l)}$ is different.

4 Results

This section presents our numerical experiments. We first study the impact of the uncertainties on the physical forcing and the observation configuration on the performance of the classical DA-based calibration scheme, and second, the relevance of a proposed learning-based approach is investigated.

4.1 Evaluation of the DA-based calibration

We display in Figure 5 the distribution of the different metrics introduced in Section 2.4 for a single-member DA-based calibration of the BGC model. For illustration purposes, we first focus on observation configuration 1/7, i.e. a daily sampling for nutrients and phytoplankton and a weekly sampling for zooplankton and detritus. We report both the metrics computed from the simulated BGC time series in terms of correlation, time shift and amplitude ratio (Fig.5a-c), and the NME metric for each model parameter (Fig.5d). Uncertainty Case-0 represents a reference case where the metrics reveal how the model can be constrained with a DA-based scheme without physical forcing uncertainties. Here, where the only errors are observation noise, the calibration is not perfect as shown by the different metrics. Despite the correlation is almost perfect for all states, a systematic shift appears for $\mathbf{Z}$ and the amplitude ratio is around $1 \pm 5e-3$ for $\mathbf{N}$, $\mathbf{P}$ and $\mathbf{D}$, with values of $\mathbf{Z}$ between 0.9 and 1.1. Fig.5d highlights the parameters that are difficult to constrain, generally associated with zooplankton, such as ρ and λ, half of which have an error above 0.1, or α and β, with 29% and 15% of error above 0.1. As expected, the uncertainty Case-1 (purple shapes in Fig.5) clearly affects the calibration performance. The correlation drops to 0.98, shifts of up to 3 days appear for $\mathbf{P}$ and $\mathbf{D}$ and amplitudes are more frequently under or over estimated for $\mathbf{N}$, $\mathbf{P}$ and $\mathbf{D}$. The amplitude ratio remains between 0.98 and 1.015 with a standard deviation much larger than for Case-0, resulting with an increase from 179% to 340% of the standard deviation of the amplitude ratio of $\mathbf{N}$, $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{D}$. The calibration performance slightly worsens again for Case-2 and Case-3 (dark blue and light blue shapes in Fig.5). The distribution of correlation Corr involves lower scores, especially for Case-2 with values of 0.96. The shift increases slightly with the highest values for Case-2. Similarly, the amplitude ratio weakly worsens with standard deviation growing to 480%, (resp. 527%, 295% and 527%) in Case-2 and 360%, (resp. 400%, 219% and 364%) in Case-3.

Fig. 5d highlights the calibration errors for the BGC parameters for observation configuration 1/7. NMEs are comprised between -0.4 and 0.4 because of the distribution of the parameters set in our experiments to $\pm 20\%$ of the reference value. Parameters ρ and λ still involve large calibration uncertainties but the calibration of the other

parameters is also sensitive to the uncertainties in the physical forcings. For instance, parameters χ, ϵ, α, β, η, φ and ζ involve a larger occurrence of relative calibration error above 0.1 of 15% in Case-1, 27% in Case-2 and 20% in Case-3 (compared to 6% in Case-0). Case-2 and Case-3 differing from their statistics (Case-2 has more amplitude in its HF variations, Case-3 has higher periods), the amplitude has more impact than the inertia in the uncertainties. Overall, we report the largest uncertainties for the parameters ρ, λ, ϵ, α and β that are linked to the zooplankton grazing, while primary production and mortality parameters are better constrained.

To complement these initial results, we report in Fig.6 the performance of the DA-based calibration for uncertainty Case-1 and the three different observation configurations. It is not surprising to see that the better the sampling of the observation the better the calibration scores. The metrics upon the state reconstruction mainly highlight the decrease of performance with no direct observations on **Z** and **D** variables (scenario 1/0). From the distribution of the correlation scores (Fig.6a), **N**, **P** and **D** seem to be equally affected by the sampling scenarios, with poorer scores above 0.97 for the configuration 1/0. The distribution of the correlation score for **Z** is even more affected, with scores dropping to 0.7. The dynamics of **N** is not sensitive to the configuration of the observation in terms of time shift compared with **P** and **D**, which have scores up to 5 days, or even **Z**, with time shifts of up to 6 days for the 1/7 and 7/7 scenarios and of up to 35 days for the 1/0 configuration. Fig.6c supports the latter results with amplitude ratios between 0.9 and 1.1 for **N**, **P** and **D**, but more distributed scores for **Z**, with values between 0.75 and 1.25 for configurations 1/7 and 7/7 and values going to 0.25 and 2.0 when there is no observation for **Z** and **D**.

When considering NME scores for the model parameters (Fig.6d), we now report a slight difference between the 1/7 and 7/7 observation configurations. Parameters ϵ, η, φ and ζ have their absolute mean errors that increase by 46%, 73% 65% and 80% from configuration 1/7 to configuration 7/7. With less information about **N** and **P**, the processes of sedimentation and remineralization and the respiration and death of **P** are likely under-constrained in the considered DA formulation. The 1/0 observation configuration has an even greater impact, with average scores increasing by around 276% for γ, 220% for β, 217% for η, 232% for φ and then 437% for ζ. In addition to the sedimentation and remineralization processes that are biased, the calibration of zooplankton grazing worsens, given that ρ, λ and α have high error scores whatever the considered observation configuration.

Assuming that the reanalysis dataset for the considered forcings involve an ensemble approach, we explore how such forcing ensembles could benefit to the targeted calibration problem. We focus on Case-3 uncertainty scenario and observation configuration 1/0. Using the ensemble approach described in Section 2.2, Figure 7 compares the calibration performance of a 100-member ensemble 4DVar method to the baseline 4DVar method that relies on a single member. Between the two schemes, all metrics depict similar patterns with a slight improvement of the ensemble mean. The parameter errors remain constant, around 0.3% between the two methods. We notice a greater relative improvement for parameters ϵ and α with a relative gain ranging from 1.6% to 1.8% for the standard deviation, as opposed to a relative gain of less than 1% for the remaining parameters. The relatively marginal improvement likely relates to the non-Gaussian distribution of the posterior of model parameters as illustrated in Figure 8. The posterior for parameters ρ, λ and α clearly depict multimodal distributions. In such cases, the ensemble mean cannot provide a relevant parameter setting. Overall, the complex multimodal posteriors highlight the impact of the noisy forcings on the relevance of the DA calibration. For a 100-member ensemble calibration, employing a single GPU configuration, an iteration is completed within 15.1 seconds, in contrast to the 0.3 seconds required by a single-member 4Dvar scheme. It is important to note that a gradient step

of 10^{-3} requires a minimum of 5000. Therefore, the enhancement brought about by enabling ensemble is not deemed to be a valuable investment.

4.2 Evaluation of the learning-based calibration

We exploit the proposed benchmarking framework to assess the performance of the proposed learning-based calibration scheme with respect to that of the DA-based scheme (single-member 4DVar) reported in the previous section. Based on the benchmarking experiment reported in the next Section, we consider the CNN scheme as our reference learning-based approach as it leads to best trade-off between calibration performance and computational complexity. Table 3 details the different scores computed according to the nine scenarios for the two considered calibration schemes. NPZD metrics highlight two main patterns in terms of inter-comparison of the DA-based and learning-based schemes. In Case-0 experiments, the DA-based approach slightly outperforms the learning-based one, whereas Case-1 with observation configurations 1/7 and 7/7 leads to very similar performance. By contrast, in all the other less favourable scenarios, we report a large improvement of the learning-based calibration scheme. The time shift metrics better reveal the differences in the calibration performance. For instance, the time shift scores for the learning-based scheme increase by 50 to 320% compared with the scores of the DA-based method in Case-0. On the contrary, in Case-2 and Case-3, the time shift scores for the learning-based scheme outperform the ones of the DA-based method, being reduced respectively by 10 and 4% with configurations 1/7, 38 and 35% with configurations 7/7 and 58, 56% with configurations 1/0. We draw similar conclusions from the model parameter metrics. Except Case-0 where the scores are equally averaged between 0.06 and 0.07, the averaged scores of the learning-based schemes remain below 0.07 for the three other cases and outperform the DA-based scheme which can lead to a poor calibration performance with averaged scores up to 0.14 given the prior knowledge that model parameters lie in a 20% interval around a reference value.

Figure 9 synthesizes our evaluation for uncertainty Case-3 and observation configuration 1/0. This experimental setup leads to the worst performance of the DA-based scheme and results in the greatest relative improvement of the learning-based approach. Fig.9a-c highlight the difficulties to retrieve accurate states from the DA-based method. Only 17% (resp. 64%) of the samples have their four correlation scores above 0.99 for the DA-based (resp. learning-based) calibration scheme. 30% are associated with time shifts below 5 days compared to 67% for the learning-based method. The amplitude ratios are between 0.8 and 1.20 for 47% of the samples with the DA-based method compared to 67% with the learning-based scheme. In this scenario, while the latter better reconstructs each state than the DA-based method, it still has difficulty finding the correct dynamics of the zooplankton. This likely relates to the selected range of parameters. As shown in Fig.4, the amplitude of zooplankton concentration can be as high as $3\mu mol.L^{-1}$ as it can remain around $10^{-2}\mu mol.L^{-1}$, which is the threshold concentration imposed in the simulation. The existence of bifurcation point in parameter space where $\mathbf{Z}$ concentrations remain very low for a given parameter range makes highly challenging the calibration problem for zooplankton processes.

Figure 9d reveals how the learning-based method improves the calibration according to the model parameter metrics. The DA-based calibration results in an average NRMSE of 0.13 with at least one parameter NME outside -0.15 and 0.15 for 99% of the samples. By comparison, the learning-based method provides NRMSEs with an average value of 0.07, particularly with only 57% of samples with one or more NMEs outside -0.15 and 0.15. We may also emphasize that some parameters still show some sensitivity, such as parameter α or β, whereas other parameters such as χ or η involve a better score with few variability.

Scenario		Method	Correlation	Shift (days)	Amplitude ratio	Parameter $NRMSE$
Sampling	Forcing uncertainty					
1/7	Case-0	DA	1.00	**0.15**	**1.00**	**0.05**
		NN	1.00	0.63	1.01	0.06
	Case-1	DA	1.00	**0.57**	**1.00**	0.06
		NN	1.00	0.70	1.01	**0.05**
	Case-2	DA	1.00	0.92	**1.00**	0.08
		NN	1.00	**0.83**	1.01	**0.06**
	Case-3	DA	1.00	0.74	**1.00**	0.07
		NN	1.00	**0.71**	1.01	**0.05**
7/7	Case-0	DA	1.00	**0.29**	**1.00**	**0.05**
		NN	1.00	0.61	1.01	0.06
	Case-1	DA	1.00	**0.69**	**1.00**	0.07
		NN	1.00	0.73	1.01	**0.06**
	Case-2	DA	1.00	1.24	0.99	0.10
		NN	1.00	**0.77**	1.01	**0.06**
	Case-3	DA	1.00	1.04	0.99	0.09
		NN	1.00	**0.68**	**1.00**	**0.06**
1/0	Case-0	DA	1.00	**0.78**	1.00	0.07
		NN	1.00	1.16	1.00	0.07
	Case-1	DA	0.99	2.21	1.01	0.13
		NN	**1.00**	**1.19**	**1.00**	**0.07**
	Case-2	DA	0.98	3.64	1.01	0.14
		NN	**0.99**	**1.54**	**1.00**	**0.07**
	Case-3	DA	0.99	2.71	1.02	0.13
		NN	**1.00**	**1.18**	**1.00**	**0.07**

Table 3. The averaged metrics upon the correlation, shift, amplitude ratio and BGC parameters $NRMSE$, defined in Section 2.4, according to a scenario. Each value represents an average over all the states or parameters and the 100 samples of the scenario.

We consider a similar analysis for the learning-based schemes as reported in Figure 10. The parameters error is distributed for the MLP (brown shapes), for the CNN (red shapes) and for the U-Net (purple shapes). The distribution are found to be highly similar across the three different neural architectures, with means within -0.05 and 0.05. Nonetheless, the MLP shows higher errors, particularly for χ, γ, η and ζ in comparison to the two other models, with standard deviations increasing by 0.02 to 0.05. On the other hand, both U-Net and CNN are equivalent in terms of BGC parameter estimation.

Finally, an experiment leveraging 100 members for the CNN test has been conducted on the Case-3 1/0 scenario. As such, the CNN is still trained on the same configuration, i.e. $N = 5000$ OSSEs with various θ, $\mathbf{U}^\star$ and $\mathbf{Y}$ realizations. However the test is performed on the $L = 100$ members of the $J = 100$ samples used for to assess the 4DVar+ensemble scheme. Once again, the ensemble scheme only provides a very slight improvement, with NRMSEs by 0.003 equal for all parameters and a standard deviation decreasing by 1% for ϵ. As the training is exactly the same for the two schemes, the computational cost remains the same between the two methods.

5 Discussion

This section discusses our main findings according to three aspects: the shortcomings of the DA-based calibration scheme with sparse and noisy observations and uncertainties on the physical forcing, the potential of deep learning schemes for ocean BGC modelling, and the generalization of our study to real-world case-studies.

5.1 Shortcomings of DA-based calibration of ocean BGC models

DA-based schemes are the state-of-the-art approaches for model calibration issues when dealing with partial and noisy observation datasets (Carrassi et al., 2018; Cheng et al., 2023). This applies to ocean biogeochemistry. Numerous studies (Lin et al., 2022; Lozano et al., 2022; Park et al., 2018; Fennel et al., 2001; McGillicuddy Jr et al., 1998) have pointed out large uncertainties and biases in BGC simulations due to the sensitivity to physical forcings, especially upper ocean dynamics and light availability. As illustrated in Park et al. (2018); Fennel et al. (2001), this translates to uncertainties and biases in the calibration of the ocean BGC model parameters. More precisely, using a variational DA scheme, they showed that uncertainties on mixed layers prevent from correctly estimating ocean BGC model parameters solely from BGC observation data. The computational complexity of ocean models, which increases with the addition of BGC processes (Brasseur et al., 2009), makes it highly-challenging to investigate ocean BGC calibration problems with operational ocean BGC setups, such as NEMO-PISCES framework (Aumont et al., 2015). In this study, we introduce a lightweight idealized 0D+t BGC modelling and observation framework. It accounts for two physical forcings, namely light availability and ocean mixing, and models four coupled components, namely nutrient, phytoplankton and zooplankton and detritus. Despite its simplicity, our results exhibit the sensitivity of the DA-based calibration of the BGC model to the uncertainties on the physical forcing coupled to a sparse observation. The non-Gaussian posteriors of the parameters issued from the ensemble DA approach underlines the non-linear nature of the considered dynamics and the impact of uncertainties on the physical forcings onto the targeted calibration problem. As pointed out for state-of-the-art BGC models in previous studies (Bagniewski et al., 2011; Friedrichs et al., 2007), we also report larger uncertainties and biases for the parameters driving the zooplankton dynamics. In particular, parameters ρ, λ, α and β are always difficult to constrain, even in the absence of forcing uncertainties (Case-0). The variability of state dynamics seems too weak to fully constrain the parameters, for example for ρ which compensates for $(1-e^{-\lambda \mathbf{P}})$. Our experiments suggest that a reduced sampling frequency for nutrients and phytoplankton affects the estimation of ϵ and η, the respiration and death rate of phytoplankton, and of the sedimentation and remineralization of detritus, φ and ζ. These are key processes in the export of particulate organic carbon (Le Moigne, 2019). In addition, when no $\mathbf{Z}$ and $\mathbf{D}$ observations are considered, we report a biased calibration of the parameters γ and β that refer to the zooplankton death and its nitrogen assimilation during grazing, to η the phytoplankton death rate, and to φ and ζ that characterize the detritus remineralization and sedimentation. Constraining remineralization in the photic layers is especially difficult because of its dynamic flux nature. As stated by Faugeras et al. (2003), observations of stocks are not sufficient to constrain fluxes. Integrating data on export or primary production in the calibration process could then help reduce estimation uncertainties.

Our idealized 0D+t BGC setup provides additional insights on the impacts of the uncertainties on physical forcings. As stated in Eq.2, the DA-based calibration relies on the minimization of a variational cost with two terms: the observation term which evaluates the goodness of the fit to the BGC observations and the prior term which constrains the reconstructed BGC dynamics to follow the calibrated BGC model (Eq.6). Importantly, this variational formulation neglects the uncertainties in the physical forcing. As illustrated in Fig.11, while the DA-based scheme actually retrieves BGC states and pa-

rameters which minimize the variational cost, it leads to a large normalized error in the estimated BGC parameters. As mentioned above, these results are in line with previous studies (Pasquier et al., 2023; Lin et al., 2022; Lozano et al., 2022; Park et al., 2018; Fennel et al., 2001; McGillicuddy Jr et al., 1998). As illustrated in our experiments, even when considering noisy but bias-free forcings (i.e., noisy forcings whose expectation is the true one), an ensemble version of the considered variational DA scheme does not reduce calibration uncertainties. Given that ocean physics reanalyses (Storto et al., 2019) also likely involve systematic biases, these results advocate to explicitly account for the uncertainties on the physical forcings in a DA calibration scheme. The classic approach to address this issue is to consider the assimilation of the coupled physics-BGC system. This is an active research topic (Goodliff et al., 2019; Berline et al., 2007). However, the implementation of such assimilation schemes remains highly-challenging both in terms of computational complexity and of numerical stability (Carrassi et al., 2018; Dowd et al., 2014), which limits their potential use in operational systems. An alternative approach would consist in parametrizing the model error covariance in Eq. 2 to account for the impact of the noisy forcings onto the BGC state dynamics. This complex task (Tandeo et al., 2018) could benefit from recent advances in bridging data assimilation and deep learning (Cheng et al., 2023; Fablet, Chapron, et al., 2021; Farchi, Bocquet, et al., 2021). Future work could extend our numerical experiments with such more advanced assimilation schemes to deliver a comprehensive benchmark of DA-based and learning-based calibration schemes for ocean BGC models within an idealized low-complexity setup.

5.2 Deep Learning for the ocean biogeochemistry

This study explores the potential learning-based schemes to overcome the shortcomings of DA-based calibration schemes for ocean BGC models. Within the considered idealised experimental setup, Fig.12 highlights the clear improvement from the learning-based approach compared to the DA-based one. With the exception of the Case-0 scenarios, which show approximately the same success rate for both methods, all the other observation and uncertainty scenarios show greater scores for the learning-based method.

When dealing with noisy forcings (i.e. Case-1/2/3 scenarios), we report on average 43% of correctly calibrated samples for the learning-based approach, by contrast to 17% for the DA-based approach. From a methodological point of view, the deep learning scheme relies on a trainable mapping from the space spanned by the observation data and the noisy physical forcings to the BGC parameter space. We train this neural mapping for a data set of more than 4000 examples combining these three data sources. Whereas the DA framework relies on the exploitation of prior knowledge to specify the considered variational cost (Eq. 2), the learning-based scheme extracts knowledge from the training dataset during the training phase. As such, the trained mapping actually accounts for both the noisy forcings and sparse observations. Importantly, at testing time, both the DA approach and the trained neural schemes are provided with the same input data, namely forcings and BGC observations, to deliver an estimate of BGC parameters. We may emphasize that the test dataset is independent on the training dataset to ensure the significance of the numerical experiments. Numerous studies in the literature support the relevance of learning-based approaches to extract relevant information from noisy input data (Gottwald & Reich, 2021; Farchi, Bocquet, et al., 2021; Brajard et al., 2021). Fig.11 also points out this key difference with the DA-based approach. While the latter minimizes variational cost (Eq.2), the deep learning scheme aims to minimize the estimation error for BGC error parameters. In turn, when computing the variational cost with the noisy physical forcings and model parameters estimated by the deep learning schemes, we report larger values of the variational cost compared with those of the DA approach. The difference increases with the scarcity of the observation. This further emphasizes that the variational cost neglecting forcing uncertainties may not provide a relevant metric to assess the quality of the estimated BGC parameters, particularly when dealing with sparse observations.

The analysis of the computational complexity of the DA-based and learning-based approaches is also an interesting aspect. When considering a learning-based scheme, most of the computational complexity lies in the training phase (Al-Jarrah et al., 2015). When dealing with OSSEs, it comprises both the creation of the training and validation datasets as well as the actual training procedure. As we state the estimation of BGC parameters as the training of a mapping from the observation and physical forcing space, we adopt a neural architecture inspired from state-of-the-art architectures used in signal and image processing (Smith et al., 2023; Z. Li et al., 2021). For such architectures and training configurations, the training datasets usually comprise from thousands to millions of samples and the training phase typically involves single-GPU setups up to a few hours. Numerous studies showed that such training procedures can scale up to significantly larger datasets and more complex neural architectures (Roussillon et al., 2023). We believe the larger constraint in terms of computational complexity lies in the creation of the training dataset using ocean BGC simulations (about 5000 in our experiments) as discussed in Section 5.3. By contrast, DA-based schemes do not involve any training phase. The experiments show that a DA-based scheme requires 5000 to 50000 iterations (according to the gradient step) of 0.3 seconds for a single member, 6.7 seconds for 20 members and 15.1 seconds for 100 members, where the training of a CNN requires 50000 iterations of 0.1 seconds. At inference time (i.e. for a trained NN), the computational complexity of variational DA schemes is thus several orders of magnitude greater than that of neural schemes, including GPU acceleration. This is particularly important for ocean BGC, as the computational cost of ocean BGC models is several times that of ocean physics models, depending on the parametrisations considered (e.g. NEMO gets 3.4 times slower with PISCES model integrated according to Maisonnave et al. (2021)). While our idealised experimental setup illustrates these differences in terms of computational complexity, recent advances in the neural emulation of atmosphere and ocean dynamics (Aouni et al., 2024; Lam et al., 2023) support this ability of neural schemes to scale up to global 3D+t dynamics with a reduced computational complexity compared to ocean BGC models.

From a methodological point of view, we leverage relatively simple neural architectures (see Section 2.3) in our numerical experiments. The focus has been on DL methods. We may emphasize that other supervised machine-learning (ML) methods such as support vector machines or random forests (Breiman, 2001; Stitson et al., 1996) could not apply directly. Such methods require designing a feature extraction step such the ML model would aim, in our case, to train a mapping between the proposed feature space and BGC model parameters. Extracting relevant features from noisy forcings and sparse observations is by itself a complex task. Our study provides an additional illustration of the widely-acknowledged ability of deep learning frameworks to train end-to-end schemes from raw data (here, noisy forcings and sparse observations) to the targeted variables (here, BGC model parameters) (LeCun et al., 2015). Numerous studies has leveraged similar end-to-end learning strategies for ocean-related applications (Febvre et al., 2023; S. A. Martin et al., 2024; Roussillon et al., 2023). The reported experiments support in this context the relevance of convolutional architecture. Despite we do not observe large differences in terms of calibration performance among the three tested architectures, we expect that more complex 1D+t or 3D+t ocean BGC settings to stress the benefit of U-Nets and other state-of-the-art neural architectures such as vision transformers (Bojesomo et al., 2023). Future work could also explore recent advances in bridging machine learning and data assimilation (Boudier et al., 2023; X. Li et al., 2022; Farchi, Laloyaux, et al., 2021; Brajard et al., 2021; Frerix et al., 2021; Fablet, Chapron, et al., 2021). Several studies support the relevance of machine learning paradigms to improve numerical models (Gupta & Lermusiaux, 2021; Farchi, Bocquet, et al., 2021; Yin et al., 2021; Bonavita & Laloyaux, 2020). Besides, end-to-end neural DA schemes (Fablet, Chapron, et al., 2021; Boudier et al., 2023) introduce and train neural architectures leveraging data assimilation formulations. While mostly explored for state estimation problems, their extensions to model calibration seems very appealing (Beauchamp et al., 2023). Addressing uncer-

tainty quantification in BGC parameter estimation is another future direction to complement our study. It could benefit from the rich literature of generative models (Goodfellow et al., 2014; Böhm et al., 2019) and probabilistic neural layers (Cheng et al., 2023).

5.3 Challenges for an application to real-world BGC data

Our study relies on simulation-only datasets, the goal being to deliver a proof-of-concept for a neural calibration of ocean BGC models. The application of the proposed methodology to real ocean BGC data naturally arises as a key question for future work. Recent studies (Febvre et al., 2023) support the relevance of neural schemes trained on simulation-only datasets to process real observations. A critical requirement for such learning-based configuration is the representativeness of the considered simulation setups. Despite its simplicity, the exploited 0D+t BGC model can provide a reduced-order representation of sea surface and upper ocean BGC dynamics (Newberger et al., 2003). By contrast, the representativeness of the considered physical forcing could be improved to match real-world datasets. Observing System Simulation Experiments (OSSEs) for ocean physics combined with operational data assimilation schemes (Fujii et al., 2019; Jean-Michel et al., 2021) provide reanalyses and reference datasets from which we could better constrain both the light availability and mixing conditions in our simulation setup. In such a 0D+t BGC configuration, we expect that the transfer to real ocean BGC datasets will not involve a significantly greater computational complexity. Such studies would also benefit from advances in AI-augmented ocean models to better account for unresolved processes and to correct model biases (Bolton & Zanna, 2019; Gupta & Lermusiaux, 2021; Farchi et al., 2023), so that the considered OSSE-based datasets better represent the real ocean dynamics.

Scaling up to more complex ocean BGC models naturally arises as a key question. Given their computational complexity, 1D+t ocean BGC configurations appear as a natural pathway for future work. Existing 1D+t ocean BGC models (Newberger et al., 2003; Schartau et al., 2001) associated with available ocean physics reanalyses and realistic OSSEs (Fujii et al., 2019; Jean-Michel et al., 2021) provide the basis to extend the proposed simulation-based learning framework with potential applications to real ocean BGC observation datasets, such as bio-ARGO float datasets (Claustre et al., 2020) and in situ mooring datasets (Gould et al., 2013). Regarding 3D+t ocean BGC dynamics, the computational burden of state-of-the-art 3D+t ocean BGC models makes the creation of a simulation-based dataset similar to that considered in this study unrealistic. However, the recent breakthroughs in the development of neural emulators for 3D+t dynamics (Aouni et al., 2024; Lam et al., 2023) could offer new possibilities to generate large-scale ensembles of ocean BGC simulations and consequently advocate for the exploration for learning-based schemes to challenging inverse problems such as model calibration.

6 Conclusion

In this study, we present a simplified ocean BGC framework to address model calibration issues. Based on a NPZD 0D+t model, 12 different scenarios were defined representing different levels of uncertainty in the physical forcings and in the observation configuration. We compared a classic model-based calibration method, using a weakly constrained 4Dvar DA formulation, to a learning-based scheme. The latter approach relies on a CNN to formulate the calibration problem as the learning of an operator mapping the observed states and the estimated physical forcings to the model parameter space.

In the absence of biased forcing, both schemes are equally effective. However, the introduction of uncertainties in light and mixing conditions highly impacts the performance of the DA-based scheme. To deliver relatively efficient BGC state predictions, assimilation leads to erroneous parameters to compensate for the biases in the physical forcings. In particular, this method has been shown to be sensitive to increasing uncertain-

ties and sparser observations, with decreasing performance, both in terms of parameter estimation and state reconstruction. Nevertheless, this method is not recommended for precise parameter recovery and BGC state forecasting, which is essential for climate studies (Sarmiento et al., 2004; Henson et al., 2022; Scheffer et al., 2001). This point is of particular significance in the context of rapid global ocean changes, emphasising the urgent need to understand the behaviour of these systems, their resilience, and to develop effective protection measures. (A. Martin et al., 2020). On the other hand, the learning-based scheme proved to be more robust under such conditions. It can retrieve relevant BGC parameter sets in most of the considered scenarios. The significant drop in its performance when no observations on zooplankton and detritus variables are available emphasizes the importance of such observations (Bagniewski et al., 2011; Faugeras et al., 2003).

The idealized framework advocates for future work dedicated for extension to real-world configurations, including real observation datasets. From a methodological standpoint, the emergence of hybrid schemes bridging machine learning, numerical modelling and data assimilation (Cheng et al., 2023; Fablet, Chapron, et al., 2021) also makes the exploration of such physics-informed and physics-constrained learning paradigms appealing. Embedding uncertainty quantification within the learning-based calibration approach is also a key challenge towards application to real ocean BGC case-studies (Cheng et al., 2023; Gupta & Lermusiaux, 2023; Garnier et al., 2016).

Open Research Statement

The authors provide the details of the Pytorch implementation and the generated data used for this study, that are available from Littaye et al. (2025).

Acknowledgments

This work was funded by the the National Centre for Scientific Research group of research OMER (Mers et Océan). It was also supported by ANR Projects Melody and OceaniX and CNES. It benefited from HPC and GPU resources from GENCI-IDRIS and CPER AIDA GPU cluster supported by The Regional Council of Brittany and FEDER. We especially thank the reviewers and the editor. Their feedback was really important and helped to make the paper clearer and stronger.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87–93.

Aouni, A. E., Gaudel, Q., Regnier, C., Van Gennip, S., Drevillon, M., Drillet, Y., & Lellouche, J.-M. (2024). Glonet: Mercator's end-to-end neural forecasting system. arXiv preprint arXiv:2412.05454.

Aumont, O., Éthé, C., Tagliabue, A., Bopp, L., & Gehlen, M. (2015). Pisces-v2: an ocean biogeochemical model for carbon and ecosystem studies. Geoscientific Model Development Discussions, 8(2), 1375–1509.

Bagniewski, W., Fennel, K., Perry, M. J., & D'asaro, E. (2011). Optimizing models of the north atlantic spring bloom using physical, chemical and bio-optical observations from a lagrangian float. Biogeosciences, 8(5), 1291–1307.

Ballarotta, M., Ubelmann, C., Pujol, M.-I., Taburet, G., Fournier, F., Legeais, J.-F., ... others (2019). On the resolutions of ocean altimetry maps. Ocean science, 15(4), 1091–1109.

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. Quarterly Journal of the Royal

Meteorological Society, 143(703), 607–633.
Beauchamp, M., Febvre, Q., Thompson, J., Georgenthum, H., & Fablet, R. (2023). Learning neural optimal interpolation models and solvers. In International conference on computational science (pp. 367–381).
Berline, L., Brankart, J.-M., Brasseur, P., Ourmières, Y., & Verron, J. (2007). Improving the physics of a coupled physical–biogeochemical model of the north atlantic through data assimilation: Impact on the ecosystem. Journal of Marine Systems, 64(1-4), 153–172.
Böhm, V., Lanusse, F., & Seljak, U. (2019). Uncertainty quantification with generative models. arXiv preprint arXiv:1910.10046.
Bojesomo, A., AlMarzouqi, H., & Liatsis, P. (2023). A novel transformer network with shifted window cross-attention for spatiotemporal weather forecasting. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 45–55.
Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. Journal of Advances in Modeling Earth Systems, 11(1), 376–399.
Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems, 12(12), e2020MS002232.
Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of compstat'2010: 19th international conference on computational statisticsparis france, august 22-27, 2010 keynote, invited and contributed papers (pp. 177–186).
Boudier, P., Fillion, A., Gratton, S., Gürol, S., & Zhang, S. (2023). Data assimilation networks. Journal of Advances in Modeling Earth Systems, 15(4), e2022MS003353.
Brady, R. X. (2017). N(utrient)-p(hytoplankton)-z(ooplankton)-d(etritus) model a toy interactive model of ocean ecosystem dynamics [ComputationalNotebook]. Retrieved from `https://github.com/bradyrx/NPZD-Model/blob/master/NPZD-Model.ipynb`
Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. Philosophical Transactions of the Royal Society A, 379(2194), 20200086.
Brasseur, P., Gruber, N., Barciela, R., Brander, K., Doron, M., Elmoussaoui, A., ... others (2009). Integrating biogeochemistry and ecology into ocean data assimilation systems. Oceanography, 22(3), 206–215.
Breiman, L. (2001). Random forests. Machine learning, 45, 5–32.
Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. Wiley Interdisciplinary Reviews: Climate Change, 9(5), e535.
Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., ... others (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. IEEE/CAA Journal of Automatica Sinica, 10(6), 1361–1387.
Circulation, O. (1989). Open university. Oceanography Course Team, Pergamon Press, Oxford.
Claustre, H., Johnson, K. S., & Takeshita, Y. (2020). Observing the global ocean with biogeochemical-argo. Annual review of marine science, 12(1), 23–48.
Claustre, H., Legendre, L., Boyd, P. W., & Levy, M. (2021). The oceans' biological carbon pumps: Framework for a research observational community approach. Frontiers in Marine Science, 8, 780052.
Doney, S. C., Lindsay, K., Caldeira, K., Campin, J.-M., Drange, H., Dutay, J.-C., ... others (2004). Evaluating global ocean carbon models: The importance of realistic physics. Global Biogeochemical Cycles, 18(3).

Dowd, M., Jones, E., & Parslow, J. (2014). A statistical overview and perspectives on data assimilation for marine biogeochemical models. Environmetrics, 25(4), 203–213.

Du, S., & Zhang, R.-H. (2024). U-net models for representing wind stress anomalies over the tropical pacific and their integrations with an intermediate coupled model for enso studies. Advances in Atmospheric Sciences, 41(7), 1403–1416.

Ducournau, A., & Fablet, R. (2016). Deep learning for ocean remote sensing: An application of convolutional neural networks for super-resolution on satellite-derived sst data. In 2016 9th iapr workshop on pattern recogniton in remote sensing (prrs) (pp. 1–6).

European Union-Copernicus Marine Service. (2019). Global ocean ensemble physics reanalysis. Mercator Ocean International. Retrieved from `https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_ENS_001_031/description` doi: 10.48670/MOI-00024

Fablet, R., Amar, M., Febvre, Q., Beauchamp, M., & Chapron, B. (2021). End-to-end physics-informed representation learning for satellite ocean remote sensing data: Applications to satellite altimetry and sea surface currents. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3, 295–302.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., & Rousseau, F. (2021). Learning variational data assimilation models and solvers. Journal of Advances in Modeling Earth Systems, 13(10), e2021MS002572.

Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., & Malartic, Q. (2021). A comparison of combined data assimilation and machine learning methods for offline and online model error correction. Journal of computational science, 55, 101468.

Farchi, A., Chrust, M., Bocquet, M., Laloyaux, P., & Bonavita, M. (2023). Online model error correction with neural networks in the incremental 4d-var framework. Journal of Advances in Modeling Earth Systems, 15(9), e2022MS003474.

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. Quarterly Journal of the Royal Meteorological Society, 147(739), 3067–3084.

Faugeras, B., Lévy, M., Mémery, L., Verron, J., Blum, J., & Charpentier, I. (2003). Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? a case study assimilating data in the northwestern mediterranean sea at the jgofs-dyfamed station. Journal of Marine Systems, 40, 99–125.

Febvre, Q., Sommer, J. L., Ubelmann, C., & Fablet, R. (2023). Training neural mapping schemes for satellite altimetry with simulation data. arXiv preprint arXiv:2309.14350.

Fennel, K., Losch, M., Schröter, J., & Wenzel, M. (2001). Testing a marine ecosystem model: sensitivity analysis and parameter optimization. Journal of Marine Systems, 28(1-2), 45–63.

Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., & Yu, L. (2022). Ocean biogeochemical modelling. Nature Reviews Methods Primers, 2(1), 76.

Fossette, S., Putman, N. F., Lohmann, K. J., Marsh, R., & Hays, G. C. (2012). A biologist's guide to assessing ocean currents: a review. Marine Ecology Progress Series, 457, 285–301.

Frerix, T., Kochkov, D., Smith, J., Cremers, D., Brenner, M., & Hoyer, S. (2021). Variational data assimilation with a learned inverse observation operator. In International conference on machine learning (pp. 3449–3458).

Friedrichs, M. A., Dusenberry, J. A., Anderson, L. A., Armstrong, R. A., Chai, F., Christian, J. R., ... others (2007). Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups.

Journal of Geophysical Research: Oceans, 112(C8).
Fujii, Y., Rémy, E., Zuo, H., Oke, P., Halliwell, G., Gasparin, F., ... others (2019). Observing system evaluation based on ocean data assimilation and prediction systems: On-going challenges and a future vision for designing and supporting ocean observational networks. Frontiers in Marine Science, 6, 417.
Garnier, F., Brankart, J.-M., Brasseur, P., & Cosme, E. (2016). Stochastic parameterizations of biogeochemical uncertainties in a 1/4° nemo/pisces model for probabilistic comparisons with ocean color data. Journal of Marine Systems, 155, 59–72.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
Goodliff, M., Bruening, T., Schwichtenberg, F., Li, X., Lindenthal, A., Lorkowski, I., & Nerger, L. (2019). Temperature assimilation into a coastal ocean-biogeochemical model: assessment of weakly and strongly coupled data assimilation. Ocean Dynamics, 69, 1217–1237.
Gottwald, G. A., & Reich, S. (2021). Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation. Physica D: Nonlinear Phenomena, 423, 132911.
Gould, J., Sloyan, B., & Visbeck, M. (2013). In situ ocean observations: A brief history, present status, and future directions. International Geophysics, 103, 59–81.
Guiavarc'h, C., Roberts-Jones, J., Harris, C., Lea, D. J., Ryan, A., & Ascione, I. (2019). Assessment of ocean analysis and forecast from an atmosphere–ocean coupled data assimilation operational system. Ocean Science, 15(5), 1307–1326.
Gupta, A., & Lermusiaux, P. F. (2021). Neural closure models for dynamical systems. Proceedings of the Royal Society A, 477(2252), 20201004.
Gupta, A., & Lermusiaux, P. F. (2023). Bayesian learning of coupled biogeochemical–physical models. Progress in Oceanography, 216, 103050.
Henson, S. A., Laufkötter, C., Leung, S., Giering, S. L., Palevsky, H. I., & Cavan, E. L. (2022). Uncertain response of ocean biological carbon export in a changing world. Nature Geoscience, 15(4), 248–254.
Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. The American Statistician, 52(2), 181–184.
IOCCG, O.-C. D. M. (2007). In w. gregg. Report of the International Ocean-Colour Coordinating Group(6).
Ismail, K., & Al-Shehhi, M. R. (2022). Reviews and syntheses: assessment of biogeochemical models in the marine environment. Biogeosciences Discussions, 1–38.
Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., ... others (2021). The copernicus global 1/12 oceanic and sea ice glorys12 reanalysis. Frontiers in Earth Science, 9, 698876.
Kane, A., Moulin, C., Thiria, S., Bopp, L., Berrada, M., Tagliabue, A., ... Badran, F. (2011). Improving the parameters of a global ocean biogeochemical model via variational assimilation of in situ data at five time series stations. Journal of Geophysical Research: Oceans, 116(C6).
Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... others (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677), 1416–1421.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436–444.
Le Moigne, F. A. (2019). Pathways of organic carbon downward transport by the oceanic biological carbon pump. Frontiers in Marine Science, 6, 634.
Lermusiaux, P., Chiu, C., Gawarkiewicz, G., Abbot, P., Robinson, A., Miller, R., ... others (2006). Quantifying uncertainties in ocean predictions in oceanogra-

phy, special issue on advances in computational oceanography, paluszkiewicz t. Harper S., Eds, 19(1), 1.

Lévy, M. (2008). The modulation of biological production by oceanic mesoscale turbulence. Transport and mixing in geophysical flows: creators of modern physics, 219–261.

Li, X., Xiao, C., Cheng, A., & Lin, H. (2022). Joint estimation of parameter and state with hybrid data assimilation and machine learning.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999–7019.

Lin, X., Massonnet, F., Fichefet, T., & Vancoppenolle, M. (2022). Impact of atmospheric forcing uncertainties on arctic and antarctic sea ice simulation in cmip6 omip. The Cryosphere Discussions, 2022, 1–40.

Littaye, J., Fablet, R., & Memery, L. (2025). Learning-based calibration of ocean carbon models to tackle physical forcing uncertainties and observation sparsity (version 6). [dataset]. Retrieved from `https://zenodo.org/doi/10.5281/zenodo.15165897`

Lozano, I. L., Sánchez-Hernández, G., Guerrero-Rascado, J. L., Alados, I., & Foyo-Moreno, I. (2022). Analysis of cloud effects on long-term global and diffuse photosynthetically active radiation at a mediterranean site. Atmospheric Research, 268, 106010.

Luo, G., Dunlap, L., Park, D. H., Holynski, A., & Darrell, T. (2023). Diffusion hyperfeatures: Searching through time and space for semantic correspondence. Advances in Neural Information Processing Systems, 36, 47500–47510.

Maisonnave, E., Berthet, S., & Séférian, R. (2021). Oasis based grid coarsening of top-pisces biogeochemistry in the nemo ocean model: performance (Unpublished doctoral dissertation). CERFACS (Toulouse).

Martin, A., Boyd, P., Buesseler, K., Cetinic, I., Claustre, H., Giering, S., ... others (2020). The oceans' twilight zone must be studied now, before it is too late. Nature, 580(7801), 26–28.

Martin, S. A., Manucharyan, G., & Klein, P. (2024). Deep learning improves global satellite observations of ocean eddy dynamics.

McGillicuddy Jr, D. J., Robinson, A., Siegel, D., Jannasch, H., Johnson, R., Dickey, T., ... Knap, A. (1998). Influence of mesoscale eddies on new production in the sargasso sea. Nature, 394(6690), 263–266.

Newberger, P. A., Allen, J. S., & Spitz, Y. (2003). Analysis and comparison of three ecosystem models. Journal of Geophysical Research: Oceans, 108(C3).

Park, J.-Y., Stock, C. A., Yang, X., Dunne, J. P., Rosati, A., John, J., & Zhang, S. (2018). Modeling global ocean biogeochemistry with physical data assimilation: A pragmatic solution to the equatorial instability. Journal of Advances in modeling earth systems, 10(3), 891–906.

Pasquier, B., Holzer, M., Chamberlain, M. A., Matear, R. J., Bindoff, N. L., & Primeau, F. W. (2023). Optimal parameters for the ocean's nutrient, carbon, and oxygen cycles compensate for circulation biases but replumb the biological pump. EGUsphere, 1–38.

Picard, T., Baker, C. A., Gula, J., Fablet, R., Mémery, L., & Lampitt, R. (2024). Estimating the variability of deep ocean particle flux collected by sediment traps using satellite data and machine learning. EGUsphere, 2024, 1–32.

Picard, T., Gula, J., Vic, C., & Mémery, L. (2024). Seasonal tracer subduction in the subpolar north atlantic driven by submesoscale fronts. Journal of Geophysical Research: Oceans, 129(9), e2023JC020782.

Pickard, G. L., & Emery, W. (1961). Descriptive physical oceanography: an introduction. Oxford: Butterworth.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and

computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18 (pp. 234–241).

Roussillon, J., Fablet, R., Gorgues, T., Drumetz, L., Littaye, J., & Martinez, E. (2023). A multi-mode convolutional neural network to reconstruct satellite-derived chlorophyll-a time series in the global ocean from physical drivers. Frontiers in Marine Science, 10, 1077623.

Royce, B., & La, S. (2011). The earth's climate sensitivity and thermal inertia. Citeseer.

Sarmiento, J. L., Slater, R., Barber, R., Bopp, L., Doney, S., Hirst, A., ... others (2004). Response of ocean ecosystems to climate warming. Global Biogeochemical Cycles, 18(3).

Schartau, M., Oschlies, A., & Willebrand, J. (2001). Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method. Deep Sea Research Part II: Topical Studies in Oceanography, 48(8-9), 1769–1800.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001). Catastrophic shifts in ecosystems. Nature, 413(6856), 591–596.

Séférian, R., Bopp, L., Gehlen, M., Orr, J. C., Ethé, C., Cadule, P., ... Madec, G. (2013). Skill assessment of three earth system models with common marine biogeochemistry. Climate Dynamics, 40, 2549–2573.

Smith, P. A., Sørensen, K. A., Buongiorno Nardelli, B., Chauhan, A., Christensen, A., St. John, M., ... Mariani, P. (2023). Reconstruction of subsurface ocean state variables using convolutional neural networks with combined satellite and in situ data. Frontiers in Marine Science, 10, 1218514.

Stitson, M., Weston, J., Gammerman, A., Vovk, V., & Vapnik, V. (1996). Theory of support vector machines. University of London, 117(827), 188–191.

Storto, A., Alvera-Azcárate, A., Balmaseda, M. A., Barth, A., Chevallier, M., Counillon, F., ... others (2019). Ocean reanalyses: recent advances and unsolved challenges. Frontiers in Marine Science, 6, 418.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., & Zhen, Y. (2018). Joint estimation of model and observation error covariance matrices in data assimilation: a review.

Trémolet, Y. (2007). Model-error estimation in 4d-var. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 133(626), 1267–1280.

Williams, R. G., & Follows, M. J. (2011). Ocean dynamics and the carbon cycle: Principles and mechanisms. Cambridge University Press.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of analysis and testing, 2(3), 249–262.

Yin, Y., Le Guen, V., Dona, J., de Bézenac, E., Ayed, I., Thome, N., & Gallinari, P. (2021). Augmenting physical models with deep networks for complex dynamics forecasting. Journal of Statistical Mechanics: Theory and Experiment, 2021(12), 124012.

Zhang, Z., Stanev, E. V., & Grayek, S. (2020). Reconstruction of the basin-wide sea-level variability in the north sea using coastal data and generative adversarial networks. Journal of Geophysical Research: Oceans, 125(12), e2020JC016402.

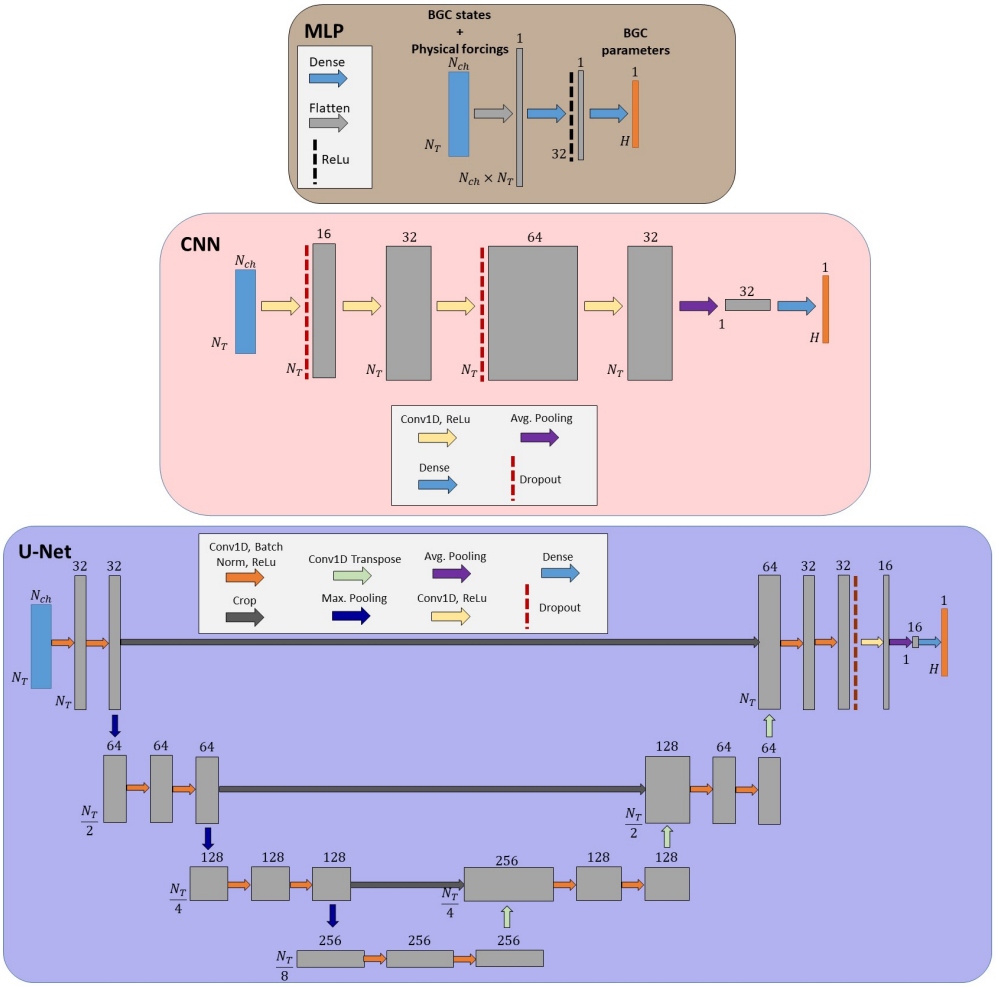

Figure 1. Architecture of three NNs used as learning-based method: A MLP composed of 2 linear layers and a ReLu layer, a CNN that contains a series of 1D-convolutional, ReLu and dropout layers and ends with pooling and dense layers, a U-Net composed of three encoder blocks (gathering maximum pooling, 1D-convolutional, batch normalisation and ReLu layers), three decoder blocks (gathering 1D-convolutional transpose layers and 1d-conv., batch norm. and ReLu layers), a dropout layer, an average pooling layer and a dense layer. Every NN receives $N_{ch} \times N_T$ size input representing observed states and physical forcings and returns H-size BGC parameter vectors as output.

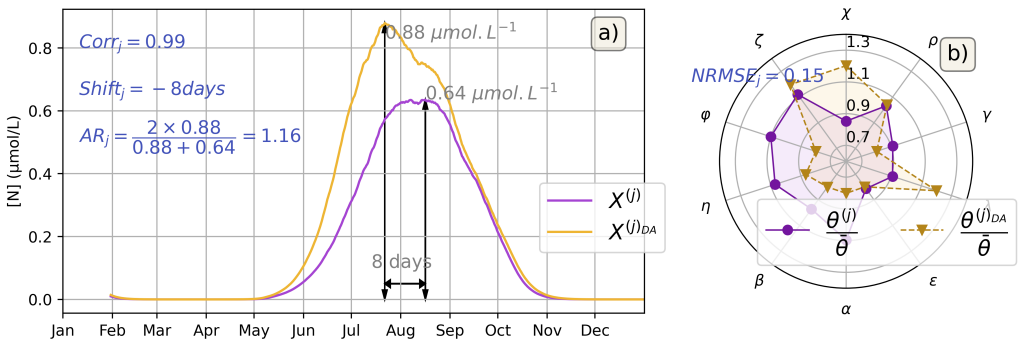

Figure 2. Illustration of the considered criteria (written in blue) to compare two BGC state time series: a) Here zooplankton time series $\mathbf{X}$ (purple) and $\mathbf{X}^{DA}$ (orange), respectively simulated from true BGC model parameters θ and estimated ones θ^{DA}. W e derive their correlation, time shift, and amplitude ratio; b), we display the normalized reference BGC parameters $\theta^{(j)}$ (purple) and the normalized estimated ones $\theta^{(j)_{DA}}$ (orange).

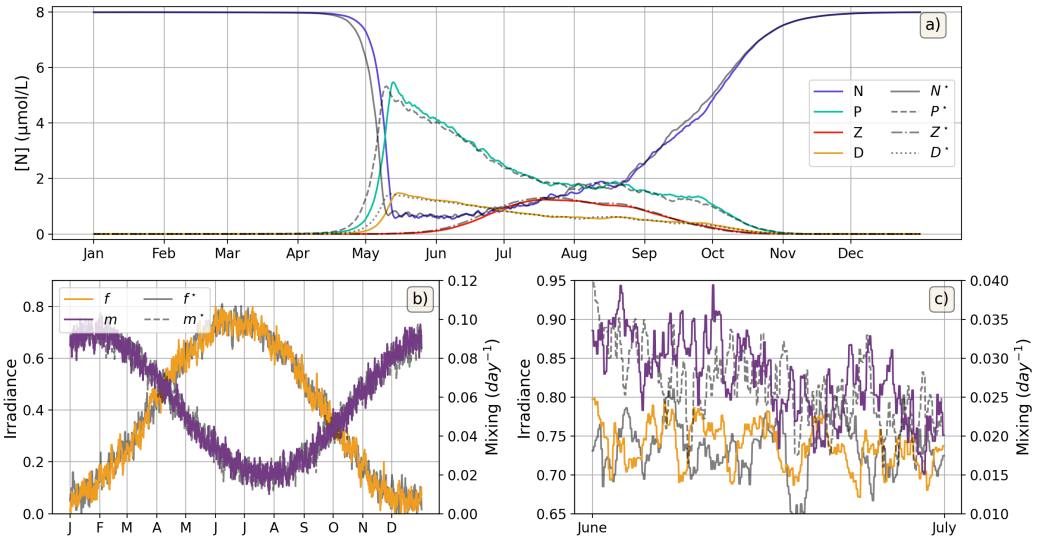

Figure 3. Visualization of the physical forcing uncertainties $\mathbf{f}^\star$ and $\mathbf{m}^\star$ upon the reconstructed states. In a), the actual ocean states ($\mathbf{N}$, $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{D}$) based on the actual physical forcings $\mathbf{U}$ and the ocean states ($\mathbf{N}^\star$, $\mathbf{P}^\star$, $\mathbf{Z}^\star$ and $\mathbf{D}^\star$) based on the re-analysis $\mathbf{U}^\star$, with b) the full variation upon a year of $\mathbf{U}$ and $\mathbf{U}^\star$, and c) their variations zoomed: on a shorter period (June to July).

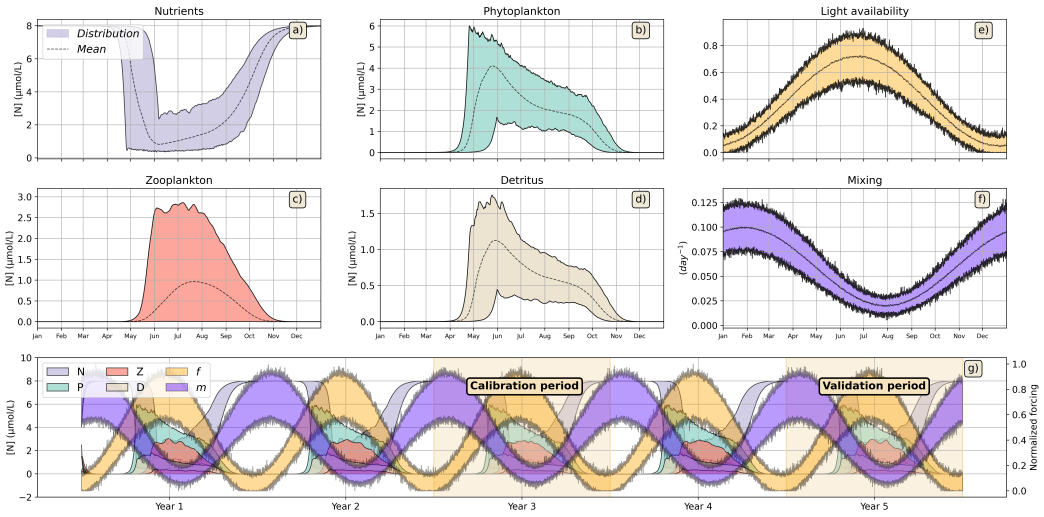

Figure 4. Distribution of the states, with shapes drew between the minimum and the maximum of all the generated data sets for the Case-1, and their associated average (dashed lines): a) nutrients, b) phytoplankton, c) zooplankton, d) detritus; and the physical forcings: e) light availability, f) mixing. g) shows the distribution of those 6 variables throughout the 5 simulated years with the training/calibration and test/validation periods.

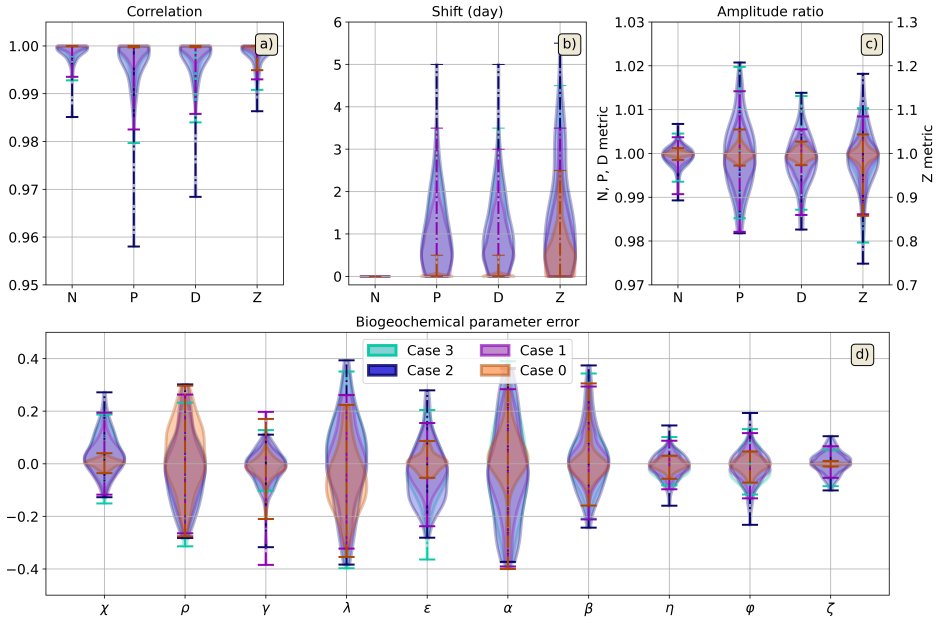

Figure 5. Violin plots of the DA-based calibration of the BGC model for the observation configuration 1/7: a) The correlation, b) the shift, c) the amplitude ratio; and d) the normalised error over each BGC parameter: χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ; for each of the 100 OSSEs. The results are related to the DA-based method for 4 levels of uncertainty upon the physical forcings: Case-0 (red), Case-1 (purple), Case-2 (dark blue) and Case-3 (light blue). The different metrics are defined in Section 2.4.

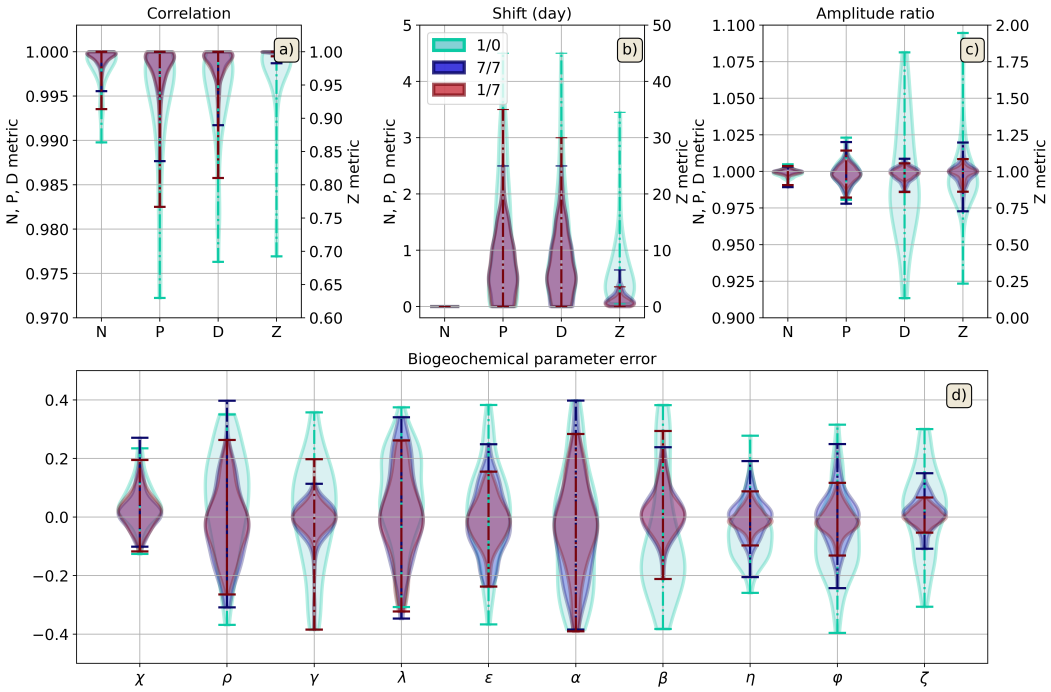

Figure 6. Violin plots of the DA-based calibration uncertainty Case-1: a) The correlation Corr, b) the shift, c) the amplitude ratio; and d) the normalised error over each BGC parameter: χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ; for each of the 100 OSSEs. The results are related to the DA-based method for 3 observation configurations: 1/7 (red), 7/7 (dark blue) and 1/0 (light blue); for the first case of uncertainty. The different metrics are defined in Section 2.4.

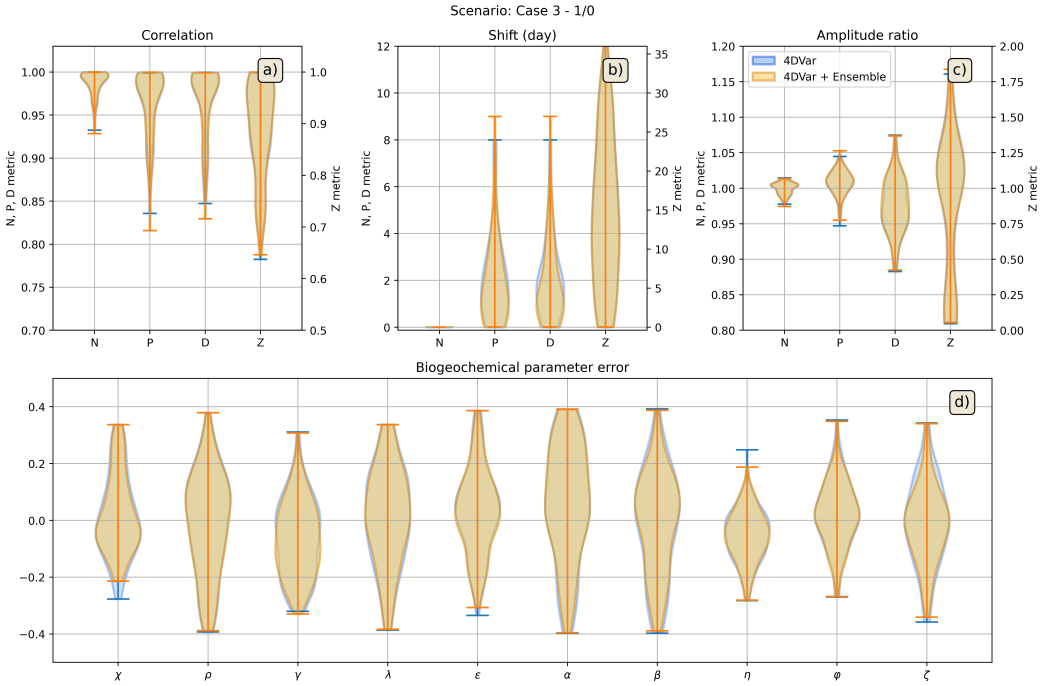

Figure 7. Violin plots of the ensemble DA-based calibration uncertainty Case-3 for observation configuration 1/0: a) The correlation, b) the shift, c) the amplitude ratio; and d) the normalised error over each BGC parameter: χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ; for each of the 100 OSSEs. The results are related to the 4DVar method enabling a single member (orange), or a 100-member ensemble (blue). The different metrics are defined in Section 2.4.

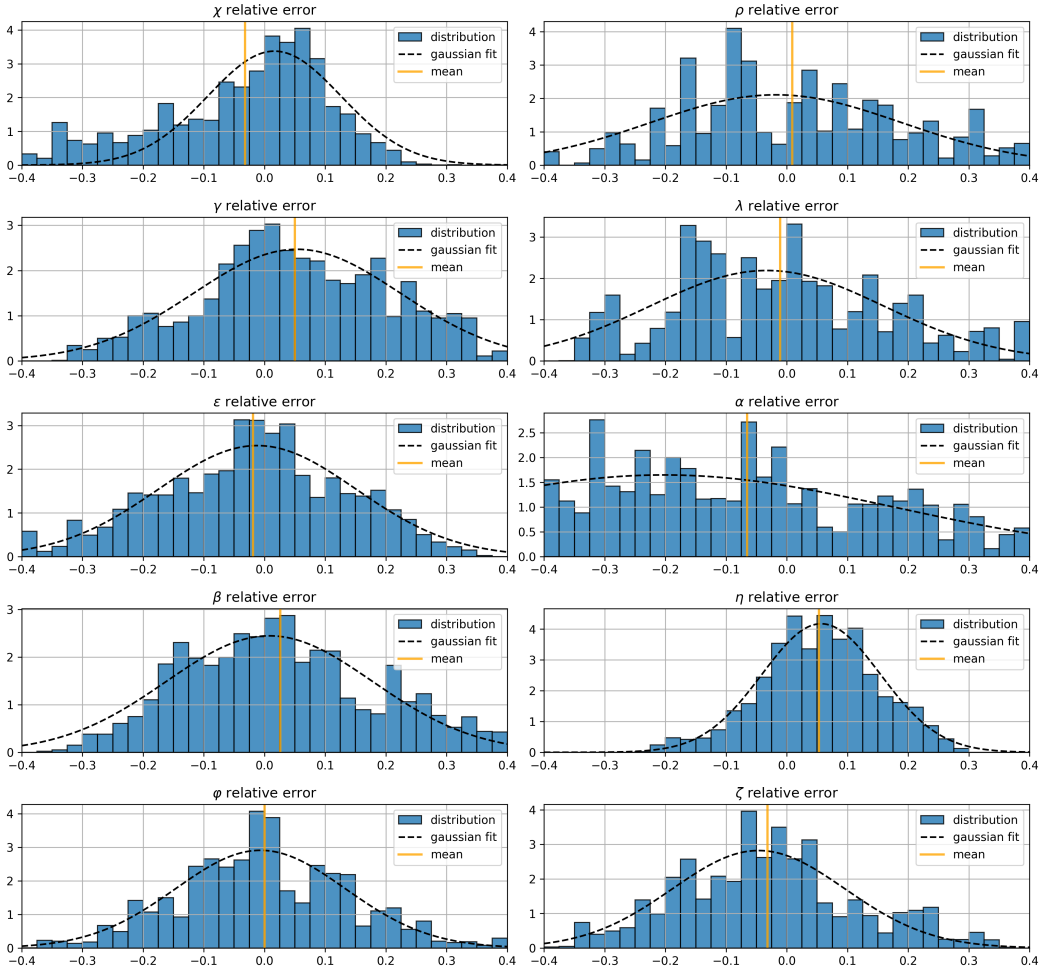

Figure 8. Parameter NMEs (defined in Section 2.4) distribution for the 4DVar method with a 100-member ensemble. The distribution is plotted for each BGC parameter: χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ; showing each of the 100 members for each of the 100 OSSEs. A Gaussian curve (black dotted line) shows tries to fit the distribution through regression over mean, standard deviation and gain. Vertical lines show the distribution mean (orange).

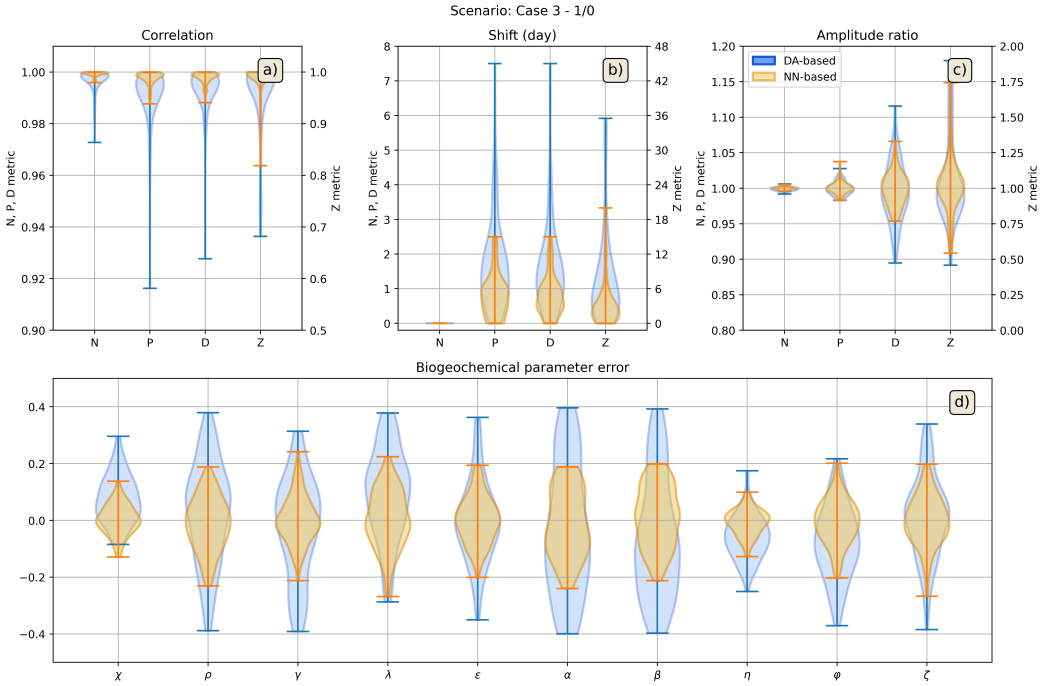

Figure 9. Violin plots of the performance metrics defined in Section 2.4 for Case-3 uncertainty scenario and 1/0 observation configuration: distribution of correlation (a), time shift (b), amplitude ratio (c) and NME (d) for all BGC parameters (χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ). We plot the distribution of the scores over the 100 samples of the test dataset for the DA-based scheme (blue) and the NN-based one (orange).

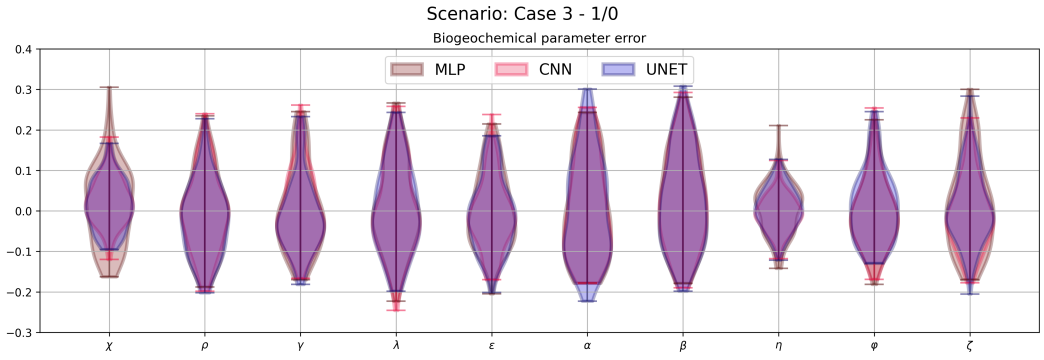

Figure 10. Violin plots of the performance metrics for Case-3 uncertainty scenario and 1/0 observation configuration: distribution of all BGC parameters NME (χ, ρ, γ, λ, ϵ, α, β, η, φ, ζ). We plot the distribution of the scores over the 100 samples of the test dataset for the MLP method (brown), the CNN method (red) and the U-Net method (purple). The metric is defined in Section 2.4.

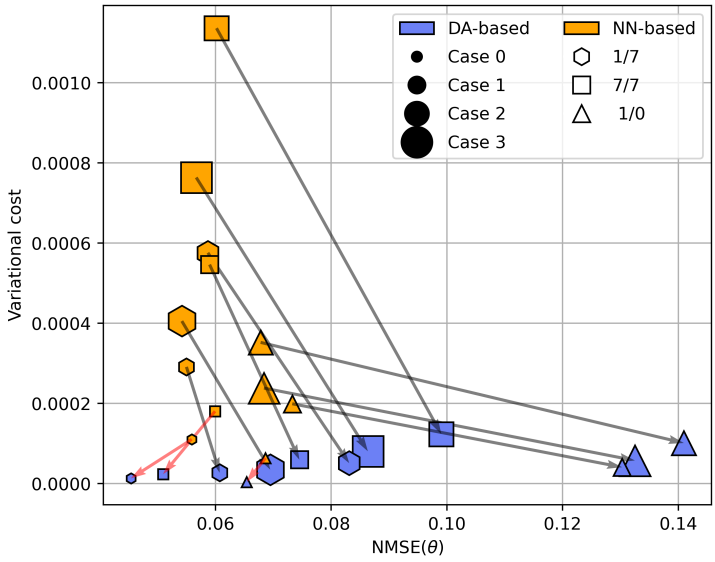

Figure 11. Scatter plot that represents the variational cost, as a function of the MSE of the parameters. The 12 scenarios corresponding to 4 different cases of uncertainties (tiny, small, medium and big-size for Case-0, 1, 2 and 3) and 3 different observation configurations (hexagons for 1/7, squares for 7/7 and triangles for 1/0) are represented by a specific marker. The blue (resp. orange) markers indicate the results from the DA-based (resp. learning-based) scheme. The black (resp. red) arrows refer to an increase (resp. a decrease) of performance from the learning-based method.

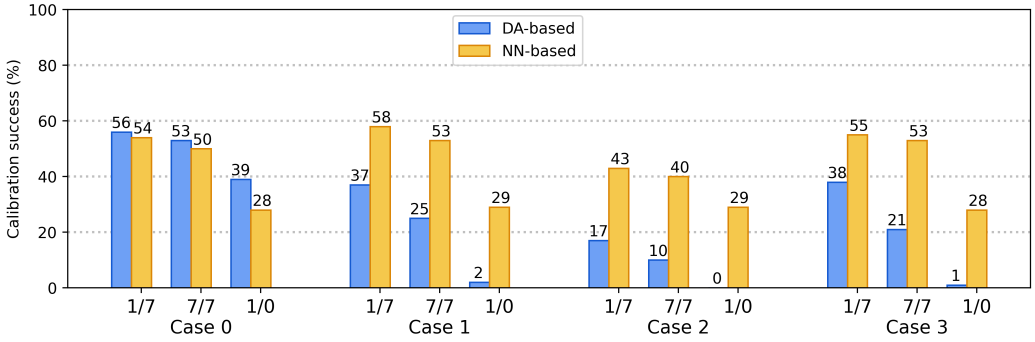

Figure 12. Percentage of samples that verify specification criteria upon the reconstructed states and the estimated parameters, i.e. correlations above 0.99, shifts below 5*days*, amplitude ratios between 0.8 and 1.2 and parameter NRMSEs below 0.15. The results are shown according to 12 scenarios (4 cases of uncertainties and 3 sampling patterns) for the DA-based calibration scheme (blue) and the learning-based one (orange).