Responses to Reviewers' Comments for Manuscript egusphere-2025-6078

Solving calibration and reanalysis challenges of ocean biogeochemical dynamics with neural schemes: a 1D vertical model case-study.

Addressed Comments for Publication to

Biogeosciences

by

Jean Littaye, Laurent Memery and Ronan Fablet

Dear Dr. Brajard,

Please find enclosed the answer for your comments regarding our previous submission entitled “Solving calibration and reanalysis challenges of ocean biogeochemical dynamics with neural schemes: a 1D vertical model case-study.”. In this document, we provide a point-by-point response to your questions. A further revised and resubmitted version would include the changes you are advising for more clarity.

Sincerely,

Jean Littaye, Laurent Memery and Ronan Fablet

Authors' Response to the Editor

General Comments. My impression is that the comparison between 4dVar and UNET is unfair. 4DVar assumes perfect forcing and imperfect model dynamics, while in the experiments, the model is actually perfect and the forcing is not. On the other hand, UNET is able to correct for the forcing error. As a consequence, it is possible that 4DVar overcorrects parameters and state variables to compensate for the forcing error, which could explain the worse results. I acknowledge that the flexibility of UNET, which does not rely on strong assumptions, is an advantage, but it would be good to include at least one fair comparison, for example, a case with no forcing error (even if forcing uncertainty is central to the work). Another option would be to rewrite the 4DVar to explicitly account for forcing error (essentially replacing the model-error term with a forcing-error term). I understand that this may be too much work for a revised version, but at the very least, the differing assumptions between the algorithms should be more clearly highlighted before the discussion section. The inconvenience of the 4DVar algorithm is acknowledged in the discussion, but it remains difficult to understand the reason for the improvement: is it the optimisation scheme itself (4DVar vs UNET) or the fact that forcing uncertainty is accounted for?

Response: Thank you for your remark.

As you note, the 4DVar approach over-corrects both the parameters and the initial state to compensate for physical uncertainty. This stems from the formulation, which does not explicitly account for noisy forcings. Whereas a classic solution is to specify a model error covariance, this is a very challenging task when dealing with uncertainties in the physical forcings. Due to the propagation of these uncertainties through the dynamical model, there is no simple analytical solution to derive a fixed model error covariance. From a model-based perspective, a more principled approach is to shift to a coupled data assimilation problem regarding the noisy forcings as a noisy observation of the true forcings and aiming at sampling the joint posterior of the forcing variables, the ocean

BGC state and the model parameters. Due to the challenges faced in the literature by coupled data assimilation systems [1, 2], we do not include this configuration as a baseline for comparison. Here, our goal is rather to emphasize the potential impact of noisy forcings, if not relevantly accounted for, on forced data assimilation configurations to address ocean BGC calibration, as it remains the classic approach used in the literature.

To deliver a fairer comparison between the DA scheme and the learning-based approach, we have carried out a complementary experiment based on noise-free forcings. In this experimental setup, the 4DVar achieves a better calibration performance than the learning-based method. These findings are in agreement with our previous study for a 0D setup [3]. They suggest the greater sensitivity of the 4DVar scheme to forcings' uncertainties representative of real ocean physics reanalysis datasets. By contrast, the neural approach learns from the training dataset how to account for noisy forcings, which results in a greater calibration robustness.

We will include this noisy-free reference experiment in a revised version of the paper to better characterize the relative sensitivity to forcings' uncertainties of the benchmarked approaches.

Comment 1

L5: The complexity of the biogeochemical processes themselves is also a major source of uncertainty.

Response:

It is accurate to assert that the BGC model error is also a significant source of uncertainty in simulations. In our case, however, we assume that model error primarily arises from poorly constrained parameters rather than from errors in the equations or parameterisation. The issue of supplementary BGC model error is referred to in the discussion and should be explored in terms of perspectives.

Comment 2

L6: ocean data assimilation of the physics or the BGC observations?

Response:

The issue under discussion pertains to the data assimilation process for BGC model calibration, regardless of the type of data employed. However, assimilating solely BGC observations is hindered by physical uncertainties, while physics only cannot accurately constrain the BGC parameters [4, 5]. Combining both sources of information provides the best results; nonetheless, this requires the implementation of a coupled physical–BGC model, which is computationally demanding when employing 3D models. A future direction is to compare the proposed learning-based calibration approach with a coupled DA scheme, under a 1D framework.

This issue will be clarified in the revised manuscript.

Comment 3

L42–55: Same remark: this could be synthesised in a few lines, reminding the reader that observations are scarce, incomplete at the surface due to clouds, lack high-resolution physics information, and are very limited in the subsurface.

Response:

This part will be synthesized in a next version.

Comment 4

L7; L80: The term BGC dynamics is not very clear in this context. While BGC processes are dynamic, the reanalysis reconstructs the state, not the dynamics themselves.

Response:

In this context, the term 'BGC dynamics' is used in a general sense to describe the BGC variable signal, rather than the parameterisation itself. In the revised version, this will instead be referred to as 'BGC variability'.

Comment 5

L110: The G term -> The G factor?

Response:

The G term is indeed a factor. This will be specified in the next version.

Comment 6

Eq. 1 and Eq. 2: These are two time-dependent equations, but it is unclear how they are linked. Are NO3, NH4, etc. defined over the vertical?

Response:

In Eq. 1, the SMS (source minus sink) term is resolved using Eq. 2, where C refers to all tracers (NO_3, NH_4, P, Z and D). Combining the two equations for nitrate concentration, for example, gives the following equation:

$$\frac{\partial \mathbf{NO_3}}{\partial t} = \frac{\partial}{\partial z}\left(\mathbf{K_z}\frac{\partial \mathbf{NO_3}}{\partial z}\right) + \mu \mathbf{NH_4} - G\frac{\mathbf{NO_3}}{\kappa + \mathbf{NO_3}} e^{-\Psi \mathbf{NH_4}} P + \lambda\left(\beta - \mathbf{NO_3}\right) \quad (1)$$

Comment 7

Eq. 3: how is U linked to Kz?

Response:

U is a multi-dimensional tensor, composed of the PAR (I_0) and the vertical diffusion coefficient (K_z). This will be specified in the revised version.

Comment 8

L203: The variables have time and depth dimensions, but in the previous paragraph (L158), U had no depth dimension.

Response:

In the previous paragraph, the depth dimension was only implied through the definition of the N_z-dimensional vector $\mathbf{U}(t, p)$, which contains the forcing at time t and horizontal location p across all depths. Now, because the observation operator $\mathcal{H}(.)$ returns measurements at specific, possibly non-uniform depths depending on the scenario, it is necessary to represent the depth dimension explicitly.

Comment 9

Eq. 5: The cost function is not introduced in the text and appears suddenly.

Response:

The equation is referenced on L237.

Comment 10

L247: It is difficult to understand that $M^{(i)}$ integrates from time 0 to time i. This could be clarified. Perhaps start with M_θ and then introduce the composition of several M_θ.

Response:

We will revised the manuscript as suggested by the reviewer.

Comment 11

L252: Equation is not numbered. There seems to be an inconsistency in the Delta notation. From line 238, Delta appears to be a time value, while in L251 it appears to be an index.

Response:

The symbol Δ refers to an index such that there are d sub-periods, with $T\delta t = d\Delta\delta t$. Therefore, Δ represents the number of time steps δt within a sub-period.

The notation will be clarified in the revised manuscript.

Comment 12

Eq. 7: The role of the factor tau_{DA} is unclear. In traditional DA, the weighting between the background term, the model-error term, and the observation term is handled by the covariance matrices B and R.

Response:

Thank you for this remark, the factor τ_{DA} can indeed be handled by the covariance matrix B.

This will be revised accordingly.

Comment 13

L268: Why is this needed since the cost is computed in the observation space?

Response:

We acknowledge the term "exact state" is misleading. The averaged value when the state is not observed refers to the initial condition of the gradient descent. The manuscript will be revised to clarify this point.

Comment 14

L299: The phrasing "no regard to optimisation" seems too strong. You could say that it is not fully optimised, since the focus is on demonstrating the potential of the method.

Response:

Thank you for your remark, this will be corrected in the next version.

Comment 15

Figure 5: I do not understand the metric for the parameter error. In section 3.4, it is said that the Normalized Mean Square Error is used (a positive value), but I see negative values in panel d.

Response:

Thank you for pointing this out. Indeed, the metric in question is defined as $Err(\theta) = \frac{\theta - \hat{\theta}}{\bar{\theta}}$, where $\bar{\theta}$ denotes the mean value of the parameter. Thus, it is not a normalised mean squared error, but rather a normalised (single) error.

Comment 16

L253 (353?): compared to a mean value of 0.99 and a minimum at 0.68: please specify that this refers to the UNET.

Response:

This will be specified in the next version.

Comment 17

L359: This definition could be moved to section 3.4.

Response:

This goes with your previous remark and will be corrected in the next version.

Comment 18

Section 4.2 and 4.3: Please remind the reader which algorithm is used.

Response:

The algorithm that has been used for these sections is the learning-based scheme. This can be specified in the next version.

Comment 19

L427: This third hybrid algorithm arrives suddenly. Why is it not introduced in the methodology section? The justification is unclear except for the improved a posteriori performance.

Response:

The novel approach presented here offers an additional way to evaluate the relevance of the corrected physical forcing produced by the UNet. A clear description of this method, along with its justification, will be included in the methodology section.

Comment 20

Figure 10: It is not clear how the standard deviation is computed. The caption says this is one sample at the beginning, and that there is an ensemble of 10 members at the end. Please explain this clearly, or recall in the main text how the ensemble is computed.

Response:

In this context, the term "sample" denotes a 10-member ensemble, relying on a single set of parameters and various forcing realisations according to Eq. 3. The standard deviation is computed from the 10 members. This will be clarified in the revised manuscript.

Comment 21

L450: It would be interesting to see improvement over non-observed variables (other variables or future variables), because in this case a cubic-spline interpolation might already give a reasonable result.

Response:

As suggested, an additional configuration has been conducted using the CTD-only observation setup, in which NH_4 and Z are not observed. This experiment is shown in Figure 1. The unobserved variables, which are only weakly reconstructed by the 4DVar method, are substantially better estimated when using the UNet and hybrid approaches. Comparing these two methods, the NMSE values for NH_4 and Z are reduced by factors of about 20 to 30, relative to Table 1. While the hybrid approach does not significantly lower the NMSE, it yields a more robust ensemble with improved representativeness of the underlying distribution. These new experiments will be incorporated in the revised manuscript.

Table 1: Table of the NMSE, standard deviation and representativeness of the reconstructed BGC states error, for the three presented schemes: a 4Dvar-only-based scheme, a UNet-only-based scheme and a hybrid scheme. The experiment is conducted under forcing uncertainty of case 1 and 10-day sampling using a CTD-only configuration. The mean error is computed over the 10 members of each sample for each BGC state. The standard deviation is computed from the 10-member ensemble of normalised reconstructed states and then averaged among all samples. The representativeness of the data is indicated by the percentage of points of the true state that are comprised within the confidence interval of the distribution of the reconstructed state.

	NMSE			Standard deviation			Distribution representativeness (%)		
BGC State	4Dvar only	UNet only	Hybrid scheme	4Dvar only	UNet only	Hybrid scheme	4Dvar only	UNet only	Hybrid scheme
NO_3	0.020	0.008	**0.004**	0.104	0.064	0.107	27.4	55.7	**65.1**
NH_4	1.617	0.070	**0.068**	0.005	0.005	0.009	4.6	41.3	**63.4**
P	0.639	0.088	**0.059**	0.061	0.056	0.063	35.4	**78.2**	73.8
Z	1.314	0.048	**0.035**	0.035	0.046	0.067	1.8	31.5	**50.7**
D	0.177	0.057	**0.030**	0.019	0.020	0.022	35.8	74.1	**76.3**

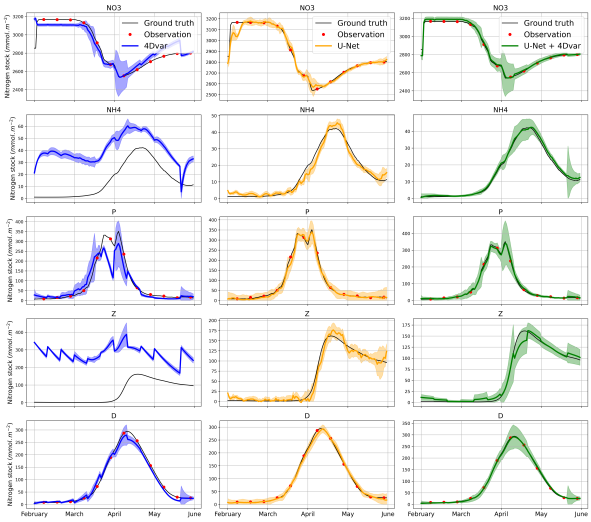

Figure 1: Reconstructed stocks associated with the five BGC states: NO_3, NH_4, P, Z and D; for one ensemble using a DA-based scheme (blue), a UNet-based scheme (orange) and a hybrid 4DVar+UNet base scheme (green), w.r.t. a scenario of case 1 forcing uncertainty, states observed with a 10-day time sampling and a **CTD-only** sampling strategy. The bold line indicates the mean stock among the 10-member ensemble, with the uncertainty, i.e., the mean value $\pm 2\sigma$ with σ being the standard deviation of the ensemble. The aforementioned confidence interval is associated with the colourized shape. The ground truth is represented by the black curve. The observed states are denoted by red dots.

Comment 22

L491: I do not see why the forcing could not be added as a control term in the 4DVar loss.

Response:

The sentence is misleading and will be corrected. In fact, it is possible to incorporate the forcing as a control term within the variational cost. However, this joint problem requires the physical forcing to be simulated online (whereas at present they are model outputs) which demands a lot more computational resources. Applying such a method to a more complex framework, such as a 3D dimensional global configuration, remains challenging for operational configurations. For the 1D framework, this has not yet been carried out because it requires a coupled model that was not available. However, it could be implemented in the future as an optimal DA reference calibration approach.

Comment 23

Section 5.2: The problem of accounting for various observation errors is not discussed. One advantage of 4DVar is the ability to handle evolving observing settings with changing observation density and varying error characteristics. Is it possible to add this in the training strategy? How would the algorithms react if observation error changes at inference time?

BGC models and observing systems evolve constantly. The training approach requires generating a large set of simulations for training data. How adaptive is the method if the model or observing system changes? Would retraining be required for every evolution?

Response:

In this context, the observation error distribution is assumed to be constant, which is a reasonable approximation for a regional case study. The observation setup is also

kept fixed for each trained/evaluated NN version. Accordingly, a separate training and evaluation has been carried out for each configuration shown in Figures 7–9. With regard to the evolving errors, they are not expected to increase, as observing systems are continuously being improved. Although applying this method to data with reduced error levels might not be fully optimal, it is not anticipated to lead to any degradation in performance.

Extending this method to a broader spatial domain naturally entails spatially varying BGC distributions and time-dependent observation errors. Within learning-based frameworks, it is a common practice to train on large sets of samples that exhibit diverse input distributions, thereby improving the model's robustness to a broad spectrum of possible densities. In addition, introducing a spatial indicator that describes the observation error density can further strengthen the method's ability to accommodate such changes [6]. Exploring how including observation error characteristics as explicit inputs influences the performance of evolving observation systems would therefore be a valuable avenue of research.

Comment 24

Section 5.3: Could you give more details on how the emulator fits in your framework? Would you have an emulator of the physics only? Could you comment on the relative computational cost of physics models versus BGC models?

Response:

Emulators provide an ensemble of ocean state estimates derived from a combination of forcing and parameters, without the necessity of solving any equation. This emerging data-driven approach offers a substantial tool to counter current ocean models that can be computationally prohibitive.

To date, the development of emulators has been primarily focused on the physical components of the ocean. [7, 8] propose realistic representation of the global ocean

circulation. These data-driven models represent a novel alternative to conventional ocean physical models. In such cases, biogeochemical models can be forced by emulated physical fields, as demonstrated in [9]. Nevertheless, the primary bottleneck is the emulation of the BGC variables. To illustrate this point, consider that the PISCES model requires 3.4 times the resources of the NEMO model. The present study is part of a series of ongoing investigations into the use of emulators in the generation of BGC components, such as surface chlorophyll and the difference in partial pressure of CO2, from physical input variables [10, 11].

References

[1] Stephen G Penny et al. "Coupled data assimilation for integrated earth system analysis and prediction: goals, challenges, and recommendations". In: (2017). URL: https://repository.library.noaa.gov/view/noaa/28431/noaa_28431_DS1.pdf.

[2] Laurence A Anderson, Allan R Robinson, and Carlos J Lozano. "Physical and biological modeling in the Gulf Stream region:: I. Data assimilation methodology". In: *Deep Sea Research Part I: Oceanographic Research Papers* 47.10 (2000), pp. 1787–1827. DOI: 10.1016/S0967-0637(00)00019-4.

[3] Jean Littaye, Ronan Fablet, and Laurent Memery. "Learning-based calibration of ocean carbon models to tackle physical forcing uncertainties and observation sparsity". In: *Journal of Advances in Modeling Earth Systems* 17.10 (2025), e2024MS004775. DOI: 10.1029/2024MS004775.

[4] Hajoon Song et al. "Data assimilation in a coupled physical-biogeochemical model of the California Current System using an incremental lognormal 4-dimensional variational approach: part 2—Joint physical and biological data assimilation twin

experiments". In: *Ocean Modelling* 106 (2016), pp. 146–158. DOI: 10.1016/j.ocemod.2016.09.003.

[5] Benoıt Pasquier et al. "Optimal parameters for the ocean's nutrient, carbon, and oxygen cycles compensate for circulation biases but replumb the biological pump". In: *EGUsphere* (2023), pp. 1–38.

[6] Joana Roussillon et al. "A Multi-Mode Convolutional Neural Network to reconstruct satellite-derived chlorophyll-a time series in the global ocean from physical drivers". In: *Frontiers in Marine Science* 10 (2023), p. 1077623. DOI: 10.3389/fmars.2023.1077623.

[7] Surya Dheeshjith et al. "Samudra: An AI global ocean emulator for climate". In: *Geophysical Research Letters* 52.10 (2025), e2024GL114318. DOI: 10.1029/2024GL114318.

[8] Anass El Aouni et al. "GLONET: Mercator's End-to-End Neural Forecasting System". In: *arXiv preprint arXiv:2412.05454* (2024).

[9] Said Ouala and Zouhair Lachkar. "Neural-BGC: An Observation-Driven Emulator for Hybrid Physical-Biogeochemical Modeling". In: (2026). DOI: 10.22541/essoar.15002003/v1.

[10] Edward Gow-Smith and Roland Séférian. *Coupling of NEMO to a neural network emulator of PISCES*. Tech. rep. Copernicus Meetings, 2026. DOI: 10.5194/egusphere-egu26-12174.

[11] Nabiz Rahpoe and Raffaele Bernardello. "A Deep Learn Emulator for Ocean Biogeochemical Modelling". In: *EGU General Assembly Conference Abstracts*. 2025, EGU25–4191. DOI: 10.5194/egusphere-egu25-4191.