Model Artificial Intelligance Ocean Observations
MAIOO

Co-directeurs : Ronan Fablet (LabSTICC, IMT Atlantique) : ronan.fablet@imt-atlantique.fr; Laurent Memery (LEMAR, CNRS) , laurent.memery@univ-brest.fr

Learning-based calibration of ocean carbon models using emerging data from new observing devices and platforms

Context. The Ocean Carbon Pump represents the processes that regulate the absorption and storage of atmospheric CO_2 in the deep ocean. This Pump plays a major role in climate and biogeochemical cycles (C, O_2, nutrients, ..). The biological part of this pump reduces unperturbed atmospheric CO_2 by 35 to 50% [1]. It is driven by photosynthesis at the ocean surface, which creates particles that are exported by gravity in the deep ocean (export production). These particles are partially remineralized in the ocean by bacterial activity and zooplankton metabolism: when they reach the deep ocean, the carbon they carry is isolated from the atmosphere for centuries.

Until the 2000s, observations of the deep ocean carbon pump were limited to a few dozen sediment traps on fixed moorings. Therefore, although these sparse data provide important first-order information in terms of vertical carbon flux, with a rough estimate of seasonal and regional variability, they are unable to strongly constrain biogeochemical models. This statement is largely based on three inescapable facts: i) these models involve a very large number of parameters, many of which are poorly known [ref]; ii) the variability of the export flux and its fate in the deep ocean is not only determined by large basin scale processes, but also by much smaller scales associated with meso and submeso scales related to ocean dynamics; iii) the vertical carbon flux is driven by a strong heterogeneity of particles, in terms of size, density, shape, composition, sinking velocity.

Since the 2010s, deep ocean observation has undergone a tremendous qualitative and quantitative evolution. In fact, new high-frequency devices on autonomous platforms are shedding light on processes that take place in the ocean depths, which have been difficult to access until now. Argo biogeochemical floats are thus deployed intensively at an increasing rate in the world ocean [2]. Thanks to technological advances in engineering sciences, critical parameters of the carbon cycle are now observable. Observations as different as oxygen, nitrate, fluorescence (chlorophyll indicator), carbon content of particles (by optical techniques), distribution of large particles and zooplankton (by imaging) are now or will soon be available as standard Argo data [ref].

Although there are still important issues to be addressed, at first order, BGC models correctly simulate the processes that govern carbon cycling in the euphotic. Nevertheless, the fate of carbon after its transfer from the surface layer remains quite coarse and highly under-parameterized [3]. The depth of remineralization is of critical importance for the carbon cycle and climate, as well as for deep ecosystem functioning [4]. Together with acoustic data, the observations obtained with Argo floats are new data, non-existent until now, that should constrain these little considered processes (more specifically particle dynamics and zooplankton distribution [4]). Given the budget and intensity of coverage expected from these

BGC Argo floats, it is crucial to quantify how these new observations can better describe the processes involved and constrain the ocean carbon models used in IPPC-type simulations.

Objectives. The main objective of this thesis is to develop a new and efficient methodology to better constrain the parameters of BGC models, in particular for biogeochemical processes in the mesopelagic layer, using these emerging observations. From a methodological point of view, model calibration problems are classically stated as minimization problems. The complexity of models has been increasing steadily over the last decades. This raises concerns about the complexity of the model/observation system in terms of the number of variables/data, the estimation of an increasing number of parameters, the quantification of uncertainties, the confidence and robustness of the model results, the nonlinear behavior of natural systems, and the cost in terms of computational power. New approaches are therefore needed. They must simplify the models while retaining the relevant processes and scales, develop efficient tools to allow exchanges between heterogeneous observations and models, and rigorously quantify uncertainties. In this regard, Artificial Intelligence has recently become an emerging field in oceanography [6], offering promising avenues both for the analysis of complex heterogeneous data and for the implementation of model-data interaction. There is a growing consensus that solutions to complex scientific problems require new methodologies that can integrate traditional physics-based modeling with Machine Learning approaches [7]. Future climate models are likely to rely on a tight integration of physical and statistical modeling paradigms, which is a fundamental change. Recent advances in AI, particularly differentiable physics, open up new avenues for addressing model parameterization issues using so-called differentiable emulators. Overall, these differentiable emulators can bridge the gap between current operational systems and AI approaches.
The proposed methodology will be applied to the NEMO/LIM3/PISCES model. This model is a national ocean/sea ice/carbon cycle dynamics model currently used in IPCC simulations and in operational oceanography (by Mercator Ocean), with a strong European dimension [5]. PISCES is considered as a model of intermediate complexity, taking into account five limiting nutrients, two types of phytoplankton, zooplankton, detritus and dissolved organic matter. Although particle dynamics and zooplankton behavior are considered in the mesopelagic layer, the processes leading to the attenuation of carbon flux with depth are nevertheless taken into account in a coarse way, as observations were lacking until now to constrain them effectively.
Proposed approach and work plan. The proposed approach will first rely on Observing System Simulation Experiments (OSSEs) to design and evaluate the proposed methodology before its application to real data. Overall, the proposed work plan includes three main tasks:

- **Task 1: OSSE design.** The PhD candidate will design an OSSE for 1D vertical model configurations for at least three contrasting target regions in the North Atlantic Ocean, Mediterranean Sea, characterized by long historical time series and intense observational coverage using new platforms and sensors: in the inter-gyre region of the Northeast Atlantic at the UK Porcupine Abyssal Plain station, in the subtropical ocean near Bermuda at the US BATS station, and at DYFAMED in the Ligurian basin. Depending on the availability of additional data in other ocean regions, these three studies could be complemented by other stations. The pseudo-observations will be simulated by the PISCES 1D model according to different model parameterizations. This pseudo-observation dataset will be representative of observations measured on BGC Argo floats, such as temperature, oxygen, fluorescence, as well as less classical

data (vertical particle size profiles, or vertical nekton migration), complemented by satellite data such as Sea Surface Temperature and surface chlorophyll.

- **Task 2: Learning-based calibration methodology.** We will exploit this OSSE context to design and evaluate the emulator approach for parameterizing BGC models from observational data. The evaluation framework will measure the influence of the following key factors: emulator architecture, sampling models, and learning criterion. With respect to emulator architecture, both direct inversion schemes and data assimilation-based architectures will be considered [8].
- **Task 3: Application to real data.** The proposed emulator-based approach will then be applied to real data in the North Atlantic, gathered at PAPS and BATS stations, and in the Mediterranean Sea (DYFAMED). Furthermore, in addition to the standard historical data, there are many observations from process cruises, which study the whole water column. Two main concrete questions will be addressed: Is it necessary to vary the parameters for different ocean regions / production regimes? Do the data require changes in the way processes are represented in PISCES? The final step of the thesis will be to adapt and extrapolate the methodology to the global ocean using the full Argo float array, supplemented by satellite data.

Added values, Synergies, Hosting teams. This project is fundamentally at the crossroads of two disciplines: ocean biogeochemistry and computer science. It contributes to the emergence of Artificial Intelligence in marine sciences. By developing innovative approaches (based on emulators), it aims to use observations (highly heterogeneous) that have not been taken into account in modeling so far, especially from the Argos BGC floats (collaboration with Laboratoire d'Océanographie de Villefranche, LOV) that are deployed worldwide at an increasing rate. Finally, we will estimate the poorly constrained parameters of a model that is intensively used in IPCC simulations and in operational oceanography (collaboration with Mercator), and with an approach capable of quantifying errors and uncertainties in a rigorous manner. The PhD student will benefit from the multidisciplinary environment developed in the framework of the AI OceaniX Chair at the crossroads of AI and oceanography (http://cia-oceanix.github.io) on the Brest campus, as well as in the Isblue University Research School (https://www.isblue.fr), which brings together the UBO, Ifremer, and the engineering schools of western Brittany. Beyond the grant, the PhD student will benefit from state-of-the-art AI resources as well as financial support for stays abroad (generally up to 6 months) in partner laboratories. These partners are Pierre Lermusiaux, MIT, Boston, a world-renowned expert in ocean data assimilation and ocean modeling using Artificial Intelligence approaches, and Adrian Martin, NOCS, UK, who has an undeniable international reputation for his expertise in the carbon pump, and in physical - biological interactions. This thesis is very strongly linked to the ANR project APERO (PI: L. Memery), aiming at constraining the fate of carbon in the mesopelagic layer on the basis of an ambitious cruise in 2023 around the PAP station. Moreover, APERO is part of JETZON, an international project supported by the UN (decade of the ocean), with Adrian Martin as PI.

Apprentissage des modèles du cycle du carbone océanique à l'aide de données émergentes provenant de nouveaux dispositifs et plates-formes d'observation

Contexte. La pompe océanique de carbone représente les processus qui régulent l'absorption et le stockage du CO2 atmosphérique dans l'océan profond. Cette pompe joue un rôle majeur dans le climat et les cycles biogéochimiques (C, O2, nutriments, ..). La partie biologique de cette pompe réduit le CO2 atmosphérique non perturbé de 35 à 50% [1]. Elle est alimentée par la photosynthèse à la surface de l'océan, qui crée des particules qui sont exportées par gravité dans l'océan profond (production d'export). Ces particules sont partiellement reminéralisées dans l'océan par l'activité bactérienne et le métabolisme du zooplancton : lorsqu'elles atteignent les profondeurs de l'océan, le carbone qu'elles transportent est isolé de l'atmosphère pendant des siècles.

Jusqu'aux années 2000, les observations de la pompe océanique de carbone étaient limitées à quelques dizaines de pièges à particules sur des mouillages fixes. Par conséquent, bien que ces données éparses fournissent des informations importantes de premier ordre en termes de flux vertical de carbone, avec une estimation approximative de la variabilité saisonnière et régionale, elles sont incapables de contraindre fortement les modèles biogéochimiques. Cette affirmation repose en grande partie sur trois faits incontournables : i) ces modèles impliquent un très grand nombre de paramètres, dont beaucoup sont mal connus [10] ; ii) la variabilité du flux d'exportation et de son devenir dans l'océan profond n'est pas seulement déterminée par des processus à l'échelle du bassin, mais aussi par des échelles beaucoup plus petites associées à des échelles méso et subméso liées à la dynamique de l'océan ; iii) le flux vertical de carbone est piloté par une forte hétérogénéité des particules, en termes de taille, de densité, de forme, de composition et de vitesse de chute.

Depuis les années 2010, l'observation de l'océan profond a connu une formidable évolution qualitative et quantitative. En effet, de nouveaux dispositifs à haute fréquence sur des plateformes autonomes permettent d'observer les processus qui se déroulent dans l'océan profond, jusqu'alors difficile d'accès. Les flotteurs biogéochimiques Argo sont ainsi déployés de manière intensive, à un rythme croissant dans l'océan mondial [2]. Grâce aux avancées technologiques des sciences de l'ingénieur, les paramètres critiques du cycle du carbone sont désormais observables. Des observations aussi différentes que l'oxygène, les nitrates, la fluorescence (indicateur de chlorophylle), le contenu en carbone des particules (par des techniques optiques), la distribution des grosses particules et du zooplancton (par imagerie) sont maintenant ou seront bientôt disponibles comme données standard Argo [2]

Bien qu'il y ait encore des questions importantes à traiter, au premier ordre, les modèles BGC simulent correctement les processus qui régissent le cycle du carbone dans l'euphotique. Néanmoins, le devenir du carbone après son transfert de la couche de surface reste assez grossier et fortement sous-paramétré [3]. La profondeur de reminéralisation est d'une importance capitale pour le cycle du carbone et le climat, ainsi que pour le fonctionnement de l'écosystème profond [4]. Avec les données acoustiques, les observations obtenues avec les flotteurs Argo sont de nouvelles données, inexistantes jusqu'à présent, qui devraient contraindre ces processus peu considérés (plus spécifiquement la dynamique des particules et la distribution du zooplancton [4]). Compte tenu du budget et de l'intensité de la couverture prévue pour ces flotteurs Argo du BGC, il est crucial de quantifier comment ces nouvelles

observations peuvent mieux décrire les processus en jeu et contraindre les modèles de carbone océanique utilisés dans les simulations de type GIEC.

Objectifs. L'objectif principal de cette thèse est de développer une méthodologie nouvelle et efficace pour mieux contraindre les paramètres des modèles BGC, en particulier pour les processus biogéochimiques en jeu dans la couche mésopélagique, en utilisant ces observations émergentes. D'un point de vue méthodologique, les problèmes de calibration de modèles sont classiquement énoncés comme des problèmes de minimisation. La complexité des modèles n'a cessé d'augmenter au cours des dernières décennies. Cela soulève des préoccupations concernant la complexité du système modèle/observation en termes de nombre de variables/données, l'estimation d'un nombre croissant de paramètres, la quantification des incertitudes, la confiance et la robustesse des résultats du modèle, le comportement non linéaire des systèmes naturels, et le coût en termes de puissance de calcul. De nouvelles approches sont donc nécessaires [9]. Elles doivent simplifier les modèles tout en conservant les processus et échelles pertinents, développer des outils efficaces pour permettre les échanges entre observations et modèles hétérogènes, et quantifier rigoureusement les incertitudes. À cet égard, l'Intelligence Artificielle est récemment devenue un domaine émergent en océanographie [6], offrant des pistes prometteuses, tant pour l'analyse de données hétérogènes complexes que pour la mise en œuvre de l'interaction modèles-données. Il y a un consensus croissant sur le fait que les solutions aux problèmes scientifiques complexes nécessitent de nouvelles méthodologies capables d'intégrer la modélisation traditionnelle basée sur la physique avec des approches Machine Learning [7]. Les futurs modèles climatiques reposeront probablement sur une intégration étroite des paradigmes de modélisation physique et statistique, ce qui constitue un changement fondamental. Les progrès récents de l'IA, et plus particulièrement de la physique différentiable, ouvrent de nouvelles voies pour aborder les questions de paramétrage des modèles à l'aide d'émulateurs dits différentiables. Globalement, ces émulateurs différentiables peuvent combler le fossé entre les systèmes opérationnels actuels et les approches d'IA**.**

La méthodologie proposée sera appliquée au modèle NEMO/LIM3/PISCES. Ce modèle est un modèle national de dynamique océanique/glace de mer/cycle du carbone actuellement utilisé dans les simulations du GIEC et en océanographie opérationnelle (par Mercator Ocean), avec une forte dimension européenne [5]. PISCES est considéré comme un modèle de complexité intermédiaire, prenant en compte cinq nutriments limitants, deux types de phytoplancton, de zooplancton, de détritus et de matière organique dissoute. Bien que la dynamique des particules et le comportement du zooplancton soient considérés dans la couche mésopélagique, les processus conduisant à l'atténuation du flux de carbone avec la profondeur sont néanmoins pris en compte de manière grossière, car les observations manquaient jusqu'à présent pour les contraindre efficacement.

Approche proposée et plan de travail. L'approche proposée s'appuiera d'abord sur des Observing System Simulation Experiments (OSSEs) pour concevoir et évaluer la méthodologie proposée avant son application à des données réelles. Globalement, le plan de travail proposé comprend trois tâches principales:

- **Tâche 1: Conception de l'OSSE.** Le candidat au doctorat concevra un OSSE pour des configurations de modèles verticaux 1D pour au moins trois régions contrastées cibles dans l'océan Atlantique Nord, en mer Méditerranée, caractérisées par de longues séries chronologiques historiques et une couverture observationnelle intense utilisant de nouvelles plateformes et de nouveaux capteurs : dans la région inter-gyre de l'Atlantique Nord-Est à la station britannique Porcupine Abyssal Plain, dans l'océan subtropical près des Bermudes à la station américaine BATS, et à DYFAMED dans le

bassin ligure. En fonction de la disponibilité de données complémentaires dans d'autres régions océaniques, ces trois études pourraient être complétées par d'autres stations. Les pseudo-observations seront simulées par le modèle PISCES 1D selon différentes paramétrisations du modèle. Ce jeu de données de pseudo-observations sera représentatif des observations mesurées sur les flotteurs BGC Argo, telles que la température, l'oxygène, la fluorescence, ainsi que des données moins classiques (profils verticaux des tailles de particules, ou migration verticale du necton), complétées par des données satellitaires telles que la température de surface et la chlorophylle de surface.

- **Tâche 2: Méthodologie de calibration basée sur l'apprentissage.** Nous exploiterons ce contexte OSSE pour concevoir et évaluer l'approche par émulateurs pour la paramétrisation de modèles BGC à partir de données d'observation. Le cadre d'évaluation mesurera l'influence des facteurs clés suivants : l'architecture des émulateurs, les modèles d'échantillonnage et le critère d'apprentissage. En ce qui concerne l'architecture des émulateurs, les schémas d'inversion directe et les architectures basées sur l'assimilation de données seront considérés [8].
- **Tâche 3: Application aux données réelles.** L'approche proposée, basée sur l'utilisation d'émulateur, sera ensuite appliquée à des données réelles dans l'Atlantique Nord, collectées aux stations PAPS et BATS, et dans la mer Méditerranée (DYFAMED). En outre, en plus des données historiques standard, il existe de nombreuses observations provenant de campagnes de processus, qui étudient l'ensemble de la colonne d'eau. Deux grandes questions concrètes seront abordées : Est-il nécessaire de faire varier les paramètres en fonction des différents régions océaniques / régimes de production ? Les données nécessitent-elles des changements dans la façon dont les processus sont représentés dans PISCES ? La dernière étape de la thèse sera d'adapter et d'extrapoler la méthodologie à l'océan global en utilisant le réseau complet de flotteurs Argo, complété par des données satellitaires.

Valeurs ajoutées, Synergies, Équipes d'accueil. Ce projet se situe fondamentalement au carrefour de deux disciplines : la biogéochimie des océans et l'informatique. Il contribue à l'émergence de l'Intelligence Artificielle dans les sciences marines. En développant des approches innovantes (basées sur des émulateurs), il vise à utiliser des observations (fortement hétérogènes) qui n'ont pas été prises en compte dans la modélisation jusqu'à présent, notamment à partir des flotteurs BGC Argos (collaboration avec Laboratoire d'Océanographie de Villefranche, LOV) qui sont déployés dans le monde entier à un rythme croissant. Enfin, il s'agira d'estimer les paramètres mal contraints d'un modèle qui est intensivement utilisé dans les simulations du GIEC et en océanographie opérationnelle (collaboration avec Mercator), de plus avec une approche capable de quantifier les erreurs et les incertitudes de manière rigoureuse. Le doctorant bénéficiera de l'environnement pluridisciplinaire développé dans le cadre de la Chaire AI OceaniX aux croisements de l'IA et de l'océanographie (http://cia-oceanix.github.io) sur le campus de Brest, ainsi que dans l'Ecole Universitaire de Recherche Isblue (https://www.isblue.fr), qui regroupe l'UBO, l'Ifremer, et les écoles d'ingénieurs de l'ouest de la Bretagne. Au-delà de la bourse, le doctorant bénéficiera de ressources de pointe en matière d'IA ainsi que d'un soutien financier pour des séjours à l'étranger (généralement jusqu'à 6 mois) dans des laboratoires partenaires. Ces partenaires sont Pierre Lermusiaux, MIT, Boston, expert mondialement reconnu en assimilation de données océaniques et en modélisation de l'océan par des approches d'Intelligence Artificielle, et Adrian Martin, NOCS, UK, qui a une réputation internationale indéniable pour

son expertise dans la pompe à carbone, et dans les interactions physiques - biologiques. Cette thèse est très fortement liée au projet ANR APERO (PI : L. Memery), visant à contraindre le devenir du carbone dans la couche mésopélagique sur la base d'une ambitieuse croisière en 2023 autour de la station PAP. De plus, APERO fait partie de JETZON, un projet international soutenu par l'ONU (décennie de l'océan), avec Adrian Martin comme PI.

References.

[1] Follows, M., and T. Oguz (2012), *The ocean carbon cycle and climate*, Springer Science & Business Media.

[2] Claustre, H., K. S. Johnson, and Y. Takeshita (2020), Observing the Global Ocean with Biogeochemical-Argo, in *Annual Review of Marine Science, Vol 12*, edited by C. A. Carlson and S. J. Giovannoni, pp. 23-48, doi:10.1146/annurev-marine-010419-010956.

[3] Le Moigne, F. A. C. (2019), Pathways of Organic Carbon Downward Transport by the Oceanic Biological Carbon Pump, *Frontiers in Marine Science*, *6*, doi:10.3389/fmars.2019.00634.

[4] Martin, A., *et al*. (2020), The oceans' twilight zone must be studied now, before it is too late, *Nature*, *580*(7801), 26-28, doi:10.1038/d41586-020-00915-7.

[5] Aumont, O., C. Ethe, A. Tagliabue, L. Bopp, and M. Gehlen (2015), PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies, *Geoscientific Model Development*, *8*(8), 2465-2513, doi:10.5194/gmd-8-2465-2015.

[6] Malde, K., *et al.* (2020). Machine intelligence and the data-driven future of marine science, J. Mar. Sci., 77(4), 1274-1285.

[7] Willard, J., *et al*. (2021). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, ArXiv: 2003.04919

[8] Fablet, R., B. Chapron, L. Drumetz, E. Mmin, O. Pannekoucke, and F. Rousseau (2020), Learning Variational Data Assimilation Models and Solvers, *Journal of Advances in Modeling Earth Systems*, doi:https://doi.org/10.1029/2021MS002572.

[9] Schartau, M., et al., (2017), Reviews and syntheses: parameter identification in marine planktonic ecosystem modelling, *Biogeosciences*, *14*(6), 1647-1701, doi:10.5194/bg-14-1647-2017

[10] Kriest, I., S. Khatiwala, and A. Oschlies (2010), Towards an assessment of simple global marine biogeochemical models of different complexity, *Progress in Oceanography*, *86*(3-4), 337-360, doi:10.1016/j.pocean.2010.05.002.