Combining artificial intelligence, Earth observations, and climate models to improve predictability of ice-biogeochemistry interactions

Grace E. Kim*, Booz Allen Hamilton
Stefanie L. Mack, British Antarctic Survey
Daniel E. Kaufman, Pacific Northwest National Laboratory
* Corresponding author: kim_grace2@bah.com

Focal Area: Predictive Modeling

We describe how artificial intelligence (AI) can be combined with state-of-the-science Earth system models to better predict future regional climate responses. To demonstrate, we describe a case study in biogeochemical interactions with sea ice.

Science Challenge

Biogeochemical models are poorly constrained for high latitude systems. Machine learning methods and edge computing can be combined with Earth system models, such as the Energy Exascale Earth System Model (E3SM), to gain insight into ice-biogeochemical interactions and improve sea ice extent prediction.

Rationale

Earth system models are powerful tools for climate prediction that explicitly define many processes and interactions. Conversely, machine learning methods use a data-driven approach that maps inputs to outputs without explicitly attempting to represent the natural, causal processes involved. We propose that Earth system models, and thus climate predictions, can be improved not only by increasing our understanding of the processes involved, but also by directly integrating machine learning (Reichstein et al., 2019) and edge computing. Specifically, machine learning is useful in cases where a given process or interaction is complex, ill-constrained, or not well understood, and there is sufficient data to train a machine learning model or neural network. Edge computing is useful when processing data at collection sources is feasible and when rapid availability of processing results is advantageous. For regional climate predictability, we suggest that detailed mechanistic representations may not always be required and certain modeled processes may be replaced with machine-learned models.

We describe a strategy of using an implicit, machine learning-powered biogeochemical model to study and predict the interaction between ice and ocean biogeochemistry. A sample implementation of this modeling framework is detailed in the following section. We recommend the implementation of software development best practices to make this work shareable, reproducible, and accessible.

Narrative

Polar physical-biogeochemical interactions determine regional dynamics and impact global biogeochemical and water cycles, through exported deep-water characteristics and ice melt rates. Microbial processes and chemical transformations in existing ocean biogeochemical models are poorly constrained (Schartau et al., 2017), particularly for high latitude systems. Consequently,

DOI: https://doi.org/10.2172/1769689
DOE BER, AI4ESP 2021 #AI4ESP1066 (https://www.ai4esp.org)

explicit models fail to fully reproduce observed biogeochemical dynamics, resulting in large errors between simulations and observations. Previous studies have tested various machine learning approaches for modeling marine biogeochemistry, outputting primary production (Mattei and Scardi, 2020), chlorophyll (Jeong et al., 2006; Sammartino et al., 2018), fluorescence (Derot et al., 2020), or algal cell densities (Xiao et al., 2017). However, these studies evaluated biogeochemical systems in specific regions and without their bi-directional interactions with physical processes.

Machine learning-enhanced climate model framework

Models of sea ice dynamics require accurate representations of surface ocean heating. In the Arctic, thawing permafrost will release ancient stores of organic carbon into adjacent waterways, reaching the downstream Arctic Ocean. The presence of colored detrital matter (CDM) affects light attenuation and changes the annual heating and temperature cycle of the surface ocean (Kim et al, 2016). Although waters in the Southern Ocean are relatively clear compared to other regions, the presence of biogenic material in sea ice changes the quantity and quality of light within the ice and in the upper ocean layer (Tilzer et al., 1994; Sarpal et al., 1995). However, phytoplankton growth and abundance in the Southern Ocean is complex and certain parameters, such as maximum growth rates and carbon to chlorophyll ratios, strongly impact model outputs but are not well constrained by available observations (Mosby and Smith, 2015; Kaufman et al., 2017; 2018). While future changes to terrestrial organic material export are known to be important for the Arctic, the impact of changing biogeochemical processes on ice dynamics in the Southern Ocean remains under-examined.

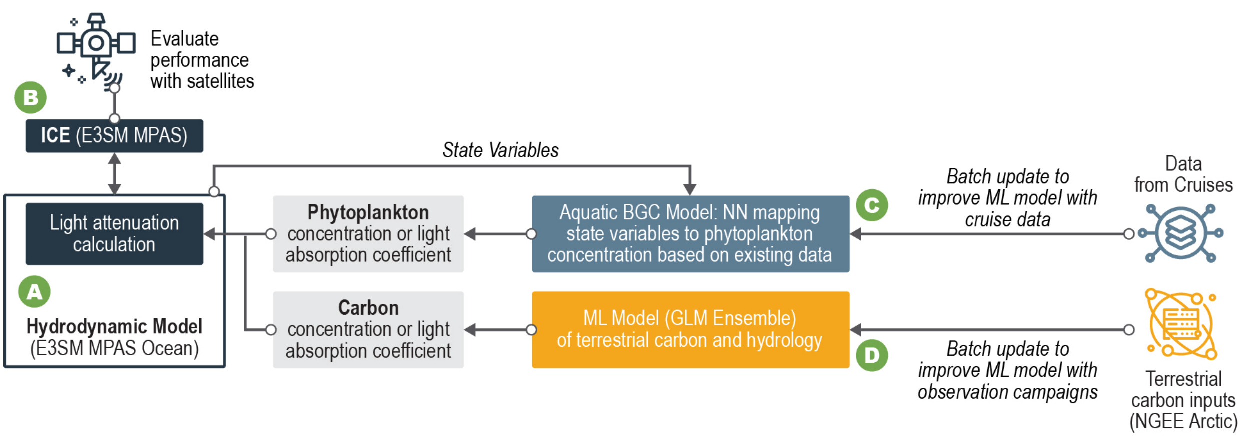

Figure 1: Schematic for a data-driven machine learning model of biogeochemistry in the DOE E3SM.

We present a framework for incorporating machine learning surrogates of marine biogeochemical-optical relationships into E3SM in Figure 1. The ocean hydrodynamic core is shown as a white box, with a light attenuation calculation that determines the irradiance vertical profile (Fig. 1; A). Radiative heating at the surface ocean affects temperature, which in turn impacts sea ice dynamics. Model performance is evaluated using the extensive satellite sea ice record (Fig. 1; B). The light attenuation calculation has two primary inputs, particulate materials and dissolved organic carbon concentrations, which are each provided by a surrogate tailored for the particular input-output mapping.

Light attenuation by phytoplankton is derived from a deep learning model with state variables from the hydrodynamic model as inputs. The model outputs proxies for phytoplankton concentration, such as chlorophyll-a or algal cell counts (Fig.1; C). Our initial expectation would be greatest success with a depth-resolved artificial neural network (such as described by Mattei and Scardi, 2020). Sufficient observational data to train a deep learning model (light blue box in Fig. 1) is already available through fluorescence-derived chlorophyll measurements from biogeochemical-argo floats, Marine mammals Exploring the Oceans Pole to pole (MEOP) programs, underwater gliders, and oceanographic cruises (e.g., Baldry et al., 2020; Chai et al., 2020); the gray box in Fig. 1 indicates where future additional data could be used to train a model that directly outputs a light absorption coefficient. Light attenuation by dissolved organic matter is derived from a machine learning model (Fig.1; D), potentially an ensemble of generalized linear models, which has been shown to be appropriate for predicting dissolved chemical species from geographic and climate variables (Kim et al., 2020). In the Arctic, inputs would include insights from the DOE's NGEE Arctic and hydrology monitoring networks, combined with the existing MOSART river model. Again, the gray box indicates a potential intermediate model of the optical contribution of colored dissolved organic matter given sufficient future data availability. For the Antarctic, this component would play a smaller role.

Edge computing and framework containerization
To maintain high levels of predictability as the system evolves, deep learning models continually require more data. An efficient and robust observational program is essential, and underwater gliders serve a unique role in such a program. Gliders can be directed post-deployment, which is particularly advantageous in harsh sampling regions such as polar seas where they may even end up under ice (Nelson et al., 2017). Through data assimilation and optimization of deployment paths, the benefits of gliders are more fully realized (Lermusiaux et al., 2017). Post-deployment optimization could be taken further by enabling gliders to serve as edge computing nodes. By running a local simulation and machine learning model subset onboard, a glider could determine and direct itself to optimal sampling locations/times, improving predictive metrics. This approach would enhance gliders' autonomy and be useful for under-ice operation, where satellite communication is not possible.

Complex software pipelines, such as the tools and dependencies required to implement the machine learning-enhanced instance of E3SM (Fig. 1), impose a strong accessibility barrier. Expertise in software development and high-performance computing (HPC) system administration is needed, as well as expertise in machine learning methods and climate science. These requirements can exclude smaller teams or individuals, particularly those with temporary employment contracts (Ph.D. students and postdocs) who may not have consistent access to HPC resources.

We encourage collaborative development within a container environment. Containers, or containerization, allow software stacks to be pre-installed in a chosen environment and deployed on any machine that has containerization software installed. Pre-built containers can be freely hosted online and allow any scientific user to access the tool or contribute to developing components of the larger model, increasing accessibility. Specific configurations of the framework can be shared by users to increase participation, enhance collaborative efforts, and make published results easily reproducible.

References

Baldry, K., Strutton, P. G., Hill, N. A., & Boyd, P. W. (2020). Subsurface chlorophyll-a maxima in the Southern Ocean. *Frontiers in Marine Science*, *7*(AUG), 1-19.

Chai, F., Johnson, K.S., Claustre, H. et al. (2020). Monitoring ocean biogeochemistry with autonomous platforms. *Nat Rev Earth Environ* 1, 315–326. https://doi.org/10.1038/s43017-020-0053-y

Derot, J., Yajima, H., Schmitt, F.G. (2020). Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas. *Ecological Informatics* 60. https://doi.org/10.1016/j.ecoinf.2020.101174

Jeong, K. S., Kim, D. K., & Joo, G. J. (2006). River phytoplankton prediction model by Artificial Neural Network: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. *Ecological Informatics*, *1*(3), 235-245.

Kaufman, D.E., Friedrichs, M.A.M., Smith, W.O., Hofmann, E.E., Dinniman, M.S., Hemmings, J.C.P. (2017). Climate change impacts on southern Ross Sea phytoplankton composition, productivity, and export. *J. Geophys. Res. Ocean.* 122, 2339–2359. https://doi.org/10.1002/2016JC012514

Kaufman, D.E., Friedrichs, M.A.M., Hemmings, J.C.P., Smith Jr., W.O. (2018). Assimilating bio-optical glider data during a phytoplankton bloom in the southern Ross Sea. *Biogeosciences* 15, 73–90. https://doi.org/10.5194/bg-15-73-2018

Kim, G. E., Gnanadesikan, A., & Pradal, M. (2016). Increased Surface Ocean Heating by Colored Detrital Matter (CDM) Linked to Greater Northern Hemisphere Ice Formation in the GFDL CM2Mc ESM, *Journal of Climate*, 29(24), 9063-9076. https://doi.org/10.1175/JCLI-D-16-0053.1

Kim, G.E., Steller, M. and Olson, S. (2020). Modeling watershed nutrient concentrations with AutoML. *Proceedings of the 10th International Conference on Climate Informatics*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3429309.3429322

Lermusiaux, P.F.J., P.J. Haley Jr., S. Jana, A. Gupta, C.S. Kulkarni, C. Mirabito, W.H. Ali, D.N. Subramani, A. Dutt, J. Lin, A.Y. Shcherbina, C.M. Lee, and A. Gangopadhyay. (2017). Optimal planning and sampling predictions for autonomous and Lagrangian platforms and sensors in the northern Arabian Sea. *Oceanography* 30(2):172–185, https://doi.org/10.5670/oceanog.2017.242.

Mattei, F., & Scardi, M. (2020). Embedding ecological knowledge into artificial neural network training: A marine phytoplankton primary production model case study. *Ecological Modelling*, 421, 108985. https://doi.org/10.1016/j.ecolmodel.2020.108985

Mosby, A., Smith, W.O. (2015). Phytoplankton Growth Rates in the Ross Sea, Antarctica. *Aquat. Microb. Ecol.* 74, 157–171. https://doi.org/10.3354/ame01733

Nelson, M., Queste, B., Smith, I., Leonard, G., Webber, B., & Hughes, K. (2017). Measurements of Ice Shelf Water beneath the front of the Ross Ice Shelf using gliders. *Annals of Glaciology*, 58(74), 41-50. doi:10.1017/aog.2017.34

Reichstein, M., Camps-Valls, G., Stevens, B. et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1

Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., Ward, B.A., Slawig, T., Oschlies, A., (2017). Reviews and syntheses: parameter identification in marine planktonic ecosystem modelling. *Biogeosciences* 14, 1647–1701. https://doi.org/10.5194/bg-14-1647-2017

Sammartino, M., Marullo, S., Santoleri, R., & Scardi, M. (2018). Modelling the vertical distribution of phytoplankton biomass in the Mediterranean Sea from satellite data: A neural network approach. *Remote Sensing*, *10*(10), 1666.

Sarpal, R.S, Mopper, K & Keiber, D., 1995. Absorbance properties of dissolved organic matter in Antarctic sea water. *Antarctic Journal*, 139–140.

Tilzer, M.M., Gieskes, W.W., Heusel, R., Fenton, N., (1994). The impact of phytoplankton on spectral water transparency in the Southern Ocean: implications for primary productivity. *Polar Biol.* 14, 127–136. https://doi.org/10.1007/BF00234975

Xiao, X., He, J., Huang, H., Miller, T. R., Christakos, G., Reichwaldt, E. S., ... & Shi, J. (2017). A novel single-parameter approach for forecasting algal blooms. *Water research*, *108*, 222-231.