Understanding Water Quality Dynamics of the Lake Water Column using Modular Compositional Learning

Bennett McAfee[a], Robert Ladwig[b], Cayelan Carey[c], Anuj Karpatne[d], Mary Lofton[c], Abhilash Neog[d], Arka Daw[d], Sophia Skoglund[a], Paul Hanson[a]

[a]Center for Limnology, University of Wisconsin–Madison, [b]Department of Ecoscience, Aarhus University, [c]Department of Biological Sciences, Virginia Polytechnic Institute and State University, [d]Department of Computer Science, Virginia Polytechnic Institute and State University

Introduction

Eco-KGML Concepts

Ecology Knowledge-Guided Machine Learning (Eco-KGML) is a paradigm that uses both machine learning (ML) and ecological knowledge to address ecological issues, including water quality.

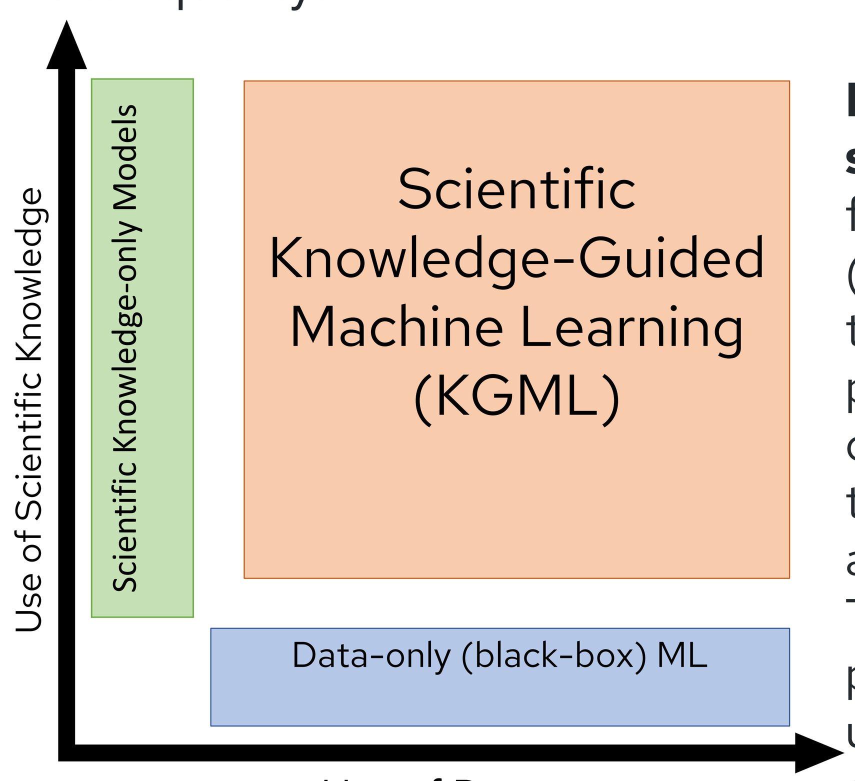


Figure 1. The operating space of KGML. Adapted from Karpatne et al. (2022)[2]. KGML utilizes both the constraints of the physical world defined by our scientific knowledge and the pattern recognition ability of machine learning. This creates more skillful predictions and greater understanding of the system.

Modular Compositional Learning

Modular Compositional Learning (MCL) is a framework for investigating ecosystem processes best modeled by either process-based (PB) or ML methods.

The overall process is segmented into sub-processes (or 'modules') that are individually modeled. Modules in MCL can be either process-based or machine learning[1].

We learn about the system from training the modules, with data informing the training process.

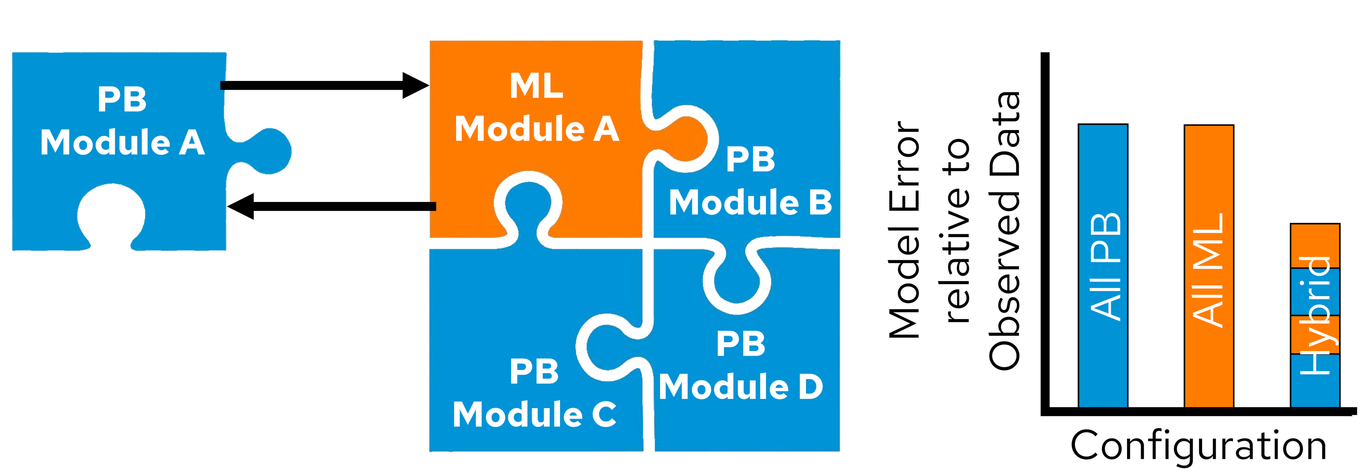


Figure 2. Example MCL framework

Aquatic Metabolism with MCL

Model Purpose and Structure

High-frequency data of water quality at depths below the surface are uncommon but important to understanding whole-lake processes.

We have created a one-dimensional model of lake metabolism and other ecosystem processes using MCL to investigate the connection between surface and at-depth water quality, as measured by water clarity, phytoplankton biomass, and dissolved oxygen concentration.

Previous findings suggest diffusive transport benefits from ML[3], the next step is to investigate other complex processes such as primary production and atmospheric gas exchange.

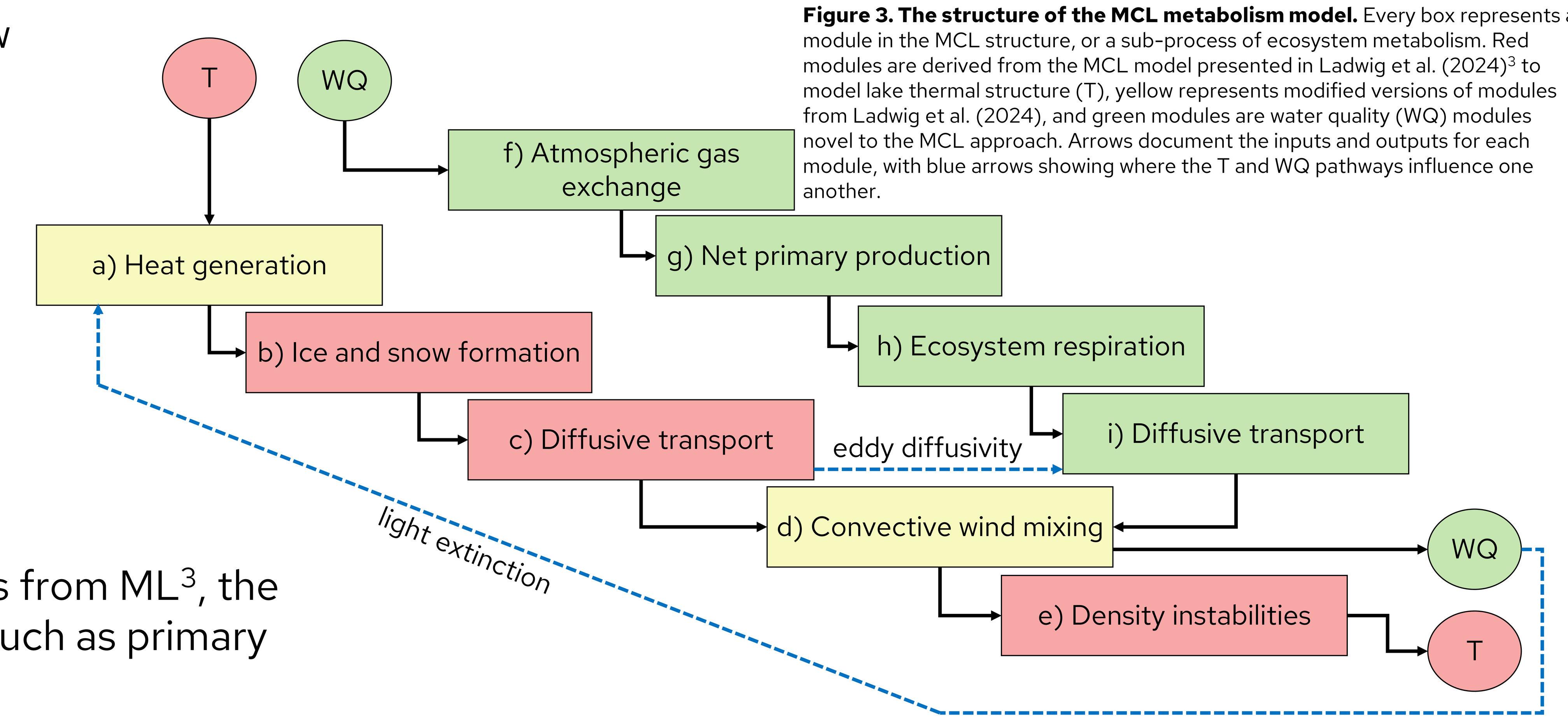
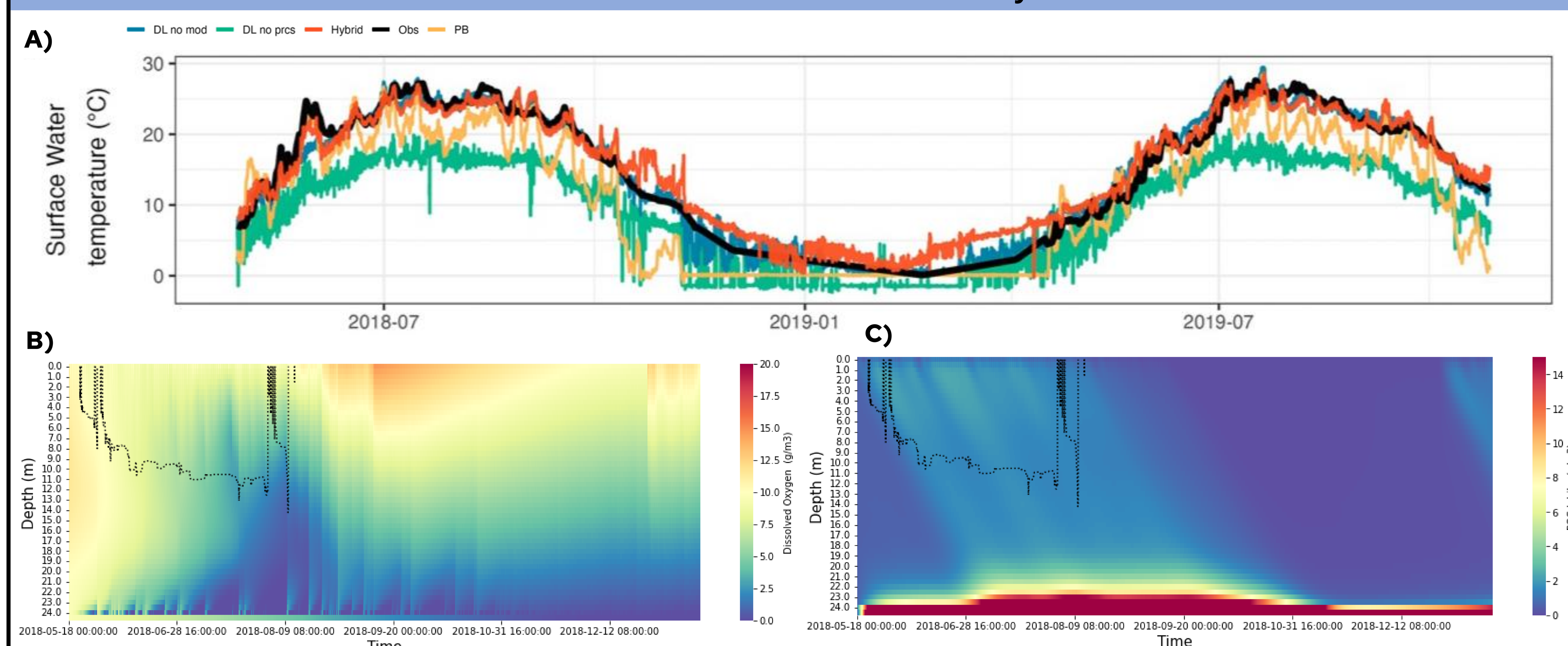


Figure 3. The structure of the MCL metabolism model. Every box represents a module in the MCL structure, or a sub-process of ecosystem metabolism. Red modules are derived from the MCL model presented in Ladwig et al. (2024)[3] to model lake thermal structure (T), yellow represents modified versions of modules from Ladwig et al. (2024), and green modules are water quality (WQ) modules novel to the MCL approach. Arrows document the inputs and outputs for each module, with blue arrows showing where the T and WQ pathways influence one another.

Preliminary Results



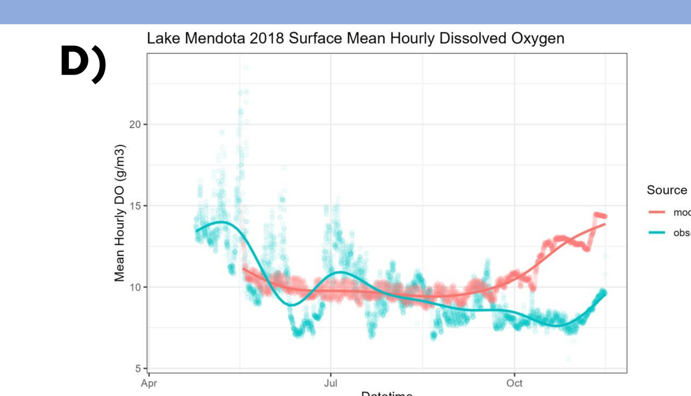


Figure 4. Modeled water quality and thermal structure of Lake Mendota (Dane County, WI). A) Adapted from Ladwig et al. (2024)[3]. Surface water temperature of Lake Mendota over time. Blue represents a pretrained deep learning model with no modularization, green represents a deep learning model with no process information, orange represents the hybrid MCL model, black represents observed data, yellow represents the process-based model. B) Modeled dissolved oxygen in Lake Mendota across all depths over one year, constructed using only process-based modules in the MCL framework. C) Modeled labile particulate organic carbon concentrations in Lake Mendota over one year, constructed using only process-based modules in the MCL framework. D) Observed dissolved oxygen (blue) and modeled dissolved oxygen (orange) at the surface of Lake Mendota over one field season. The modeled data was constructed using only the process-based modules in the MCL framework.

Early results of a purely process-based MCL model show a recreation of seasonal patterns of lake water quality. Further adjustment of the process-based modules' parameters is required to better fit the model output to observed values. Future ML integration into the MCL framework may lead to changes in model output across timescales.

Our next step is to analyze the relationship between water quality characteristics at the surface and deeper in the water column at a high frequency. Knowledge of this relationship will increase our understanding of the processes that drive vertical variation in lake water quality.

References

1. Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168
2. Karpatne, A., Kannan, R., & Kumar, V. (2022). *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. CRC Press.
3. Ladwig, R., Daw, A., Albright, E. A., Buelo, C., Karpatne, A., Meyer, M. F., Neog, A., Hanson, P. C., & Dugan, H. A. (2024). Modular Compositional Learning Improves 1D Hydrodynamic Lake Model Performance by Merging Process-Based Modeling With Deep Learning. *Journal of Advances in Modeling Earth Systems*, *16*(1), e2023MS003953. https://doi.org/10.1029/2023MS003953

Funding

This project is supported in part by grants from the US National Science Foundation (#2025982, #2213549, #2223103)

Collaborators

Center for Limnology University of Wisconsin–Madison
VIRGINIA TECH
NTL LTER
neon National Ecological Observatory Network
EPA
USGS science for a changing world

Contact and More Information

Contact the author at **bmcafee@wisc.edu**
Find more information about the project at **eco-kgml.github.io**