**Title: Upscaling global wetland methane emissions with causality guided machine learning**

**Authors/Affiliations:**
Kunxiaojia Yuan[1], Qing Zhu[1], William Riley[1], Gavin McNicol[2], Fa Li[3], Min Chen[3]
[1]Lawrence Berkeley National Laboratory, Climate & Ecosystem Sciences Division
[2]University of Illinois at Chicago, Department of Earth and Environmental Sciences
[3]University of Wisconsin-Madison Madison, Department of Forest and Wildlife Ecology

**Focal Area(s):**

- Key uncertainties and knowledge gaps where new methodology, infrastructure, or technology can advance predictive understanding of the methane cycle.
- The importance of high potential datasets or how the combination of data across spatial or temporal scales or scientific domains may lead to new scientific insights.

**Science or Technological Challenge:**

Wetland methane ($CH_4$) emissions involve many nonlinear and asynchronous processes, which can be affected by multiple environmental and biological factors. Despite promising performance demonstrated by traditional machine learning (ML) models, confounding variables often confuse traditional correlation-based ML models to mis-capture dominant drivers, thus, leading to large uncertainties in model extrapolation.

Due to the complex nature of wetland methane, the magnitude of $CH_4$ emissions as well as its responses to environmental and biological factors have shown large spatial heterogeneous characteristics, implying that extensive site observations are needed to constrain the upscaling models and thus yield reliable gridded $CH_4$ estimations. However, the current data-driven ML based wetland $CH_4$ emission products are limited by data availability, especially in high emission areas (such as tropical area, and wetland hot spots in the boreal arctic area).

Therefore, a more advanced ML model that can be robustly trained by existing dataset and more in-situ observations are urgently needed to generate a reliable global wetland methane emission dataset. Such upscaled dataset can be used for benchmarking the bottom-up biogeochemistry and top-down atmospheric inversion models, and also can be used to analyze the long-term trend and variations of wetland emissions across different regions in the world.

**Rationale:**

Methane is one of the most important global warming contributors after carbon dioxide ($CO_2$), with a Global Warming Potential (GWP) of 28-34 times of $CO_2$ over a 100-year time horizon (IPCC, 2013). Wetlands are the largest natural source of global methane, contributing 20-30% to global methane and remaining the most uncertain natural $CH_4$ source to the atmosphere (Saunois et al. 2020). Due to the limited understanding of wetland $CH_4$ emission processes and lack of observations to constrain models, large discrepancies still exist among bottom-up models and top-down models (Saunois et al. 2020). In addition, there is no widely accepted global benchmarking data product for wetland $CH_4$ emissions to evaluate, parameterize, and improve both bottom-up and top-down models. Hence, a reliable data-driven benchmark dataset of global wetland $CH_4$ emissions is urgently needed.

Data-driven ML based gridded $CH_4$ emission datasets upscaled from in situ observations play an increasingly important role on benchmarking bottom-up and top-down models. However, most current used ML models for $CH_4$ upscaling ignore the long-term dependences (between $CH_4$ emission and its drivers), and such correlation-based ML models may misidentify dominant drivers with wrong processes. Besides, lack of observation constraint, especially in high-emission areas, results in considerable uncertainties. Therefore, improvement of current ML upscaling models and collection of sufficient multi-sourced observations are both needed to generate a reliable global wetland $CH_4$ upscaling dataset.

**Narrative:**

Our objective is to generate a global wetland $CH_4$ flux emission dataset using a causality guided ML model. To achieve this, we will compile a comprehensive wetland $CH_4$ emission observation dataset with ~140 and ~180 site years of eddy covariance and chamber measurements, which will broadly cover both hotspot and non-hotspot regions across the world. Then, a physically interpretable and causality guided machine learning (causal-ML) model will be built based on our previous work (Yuan et al., 2022), which indicated that our causal-ML model can correctly capture the causal relationships between $CH_4$ emission and its drivers and achieve high prediction accuracy. Using the upscaled dataset, we will benchmark the performance of bottom-up and top-down models which participated recent global carbon project (GCP), and further investigate the predominant drivers which regulate the long-term trend and variability of wetland $CH_4$ emissions.

**References:**

IPCC, 2013. Chapter 6: Carbon and Other Biogeochemical Cycles, Climate Change 2013 The Physical Science Basis. Cambridge University Press, Cambridge.

Saunois, M. et al., 2020. The global methane budget 2000–2017. Earth System Science Data, 12(3): 1561-1623.

Yuan, K., Zhu, Q., Li, F., Riley, W.J., Torn, M., Chu, H., McNicol, G., Chen, M., Knox, S., Delwiche, K. and Wu, H., 2022. Causality guided machine learning model on wetland CH4 emissions across global wetlands. Agricultural and Forest Meteorology, 324, p.109115.