Harmonizing Natural Methane Datasets using Knowledge Guided Machine Learning
Short title: AI for Natural Methane

Youmi Oh, University of Colorado Boulder, youmi.oh@noaa.gov
Sparkle Malone, Yale University, sparkle.malone@yale.edu
Gavin McNicol, University of Illinois Chicago, gmcnicol@uic.edu
Licheng Liu, University of Minnesota, lichengl@umn.edu

Project Summary

This synthesis working group will harmonize multi-source datasets using a knowledge-guided machine learning (KGML) framework to estimate global natural methane (CH_4) fluxes. The magnitude and long-term trends of global CH_4 emissions from wetlands and consumptions from soil sinks are highly uncertain due to lack of harmonized measurements and overlooked biogeochemical processes. Accurately quantifying global natural CH_4 fluxes is extremely important to reduce biases in current and future global CH_4 budgets. This working Group will improve our understanding of natural CH_4 budgets by harmonizing simulated and observed flux datasets from global wetlands and soil sinks. To synthesize these datasets, we propose to use KGML by seamlessly integrating advanced top-down and bottom-up models, machine learning techniques, and multi-source datasets. As an output of this working group, we will generate and publicly share harmonized measurement datasets and monthly global natural CH_4 flux products from KGML at 0.5-degree resolution from 1980 to present.

Public Summary

Atmospheric methane (CH_4) is the second most powerful greenhouse gas after carbon dioxide and grew at the fastest rate ever recorded in 2020-2022. Slowing or reversing the accelerating growth in atmospheric CH_4 will require an improved understanding of the global CH_4 budget, which is currently underconstrained. Natural CH_4 budgets are responsible for ~40% of the total global CH_4 budgets but remain the most uncertain factor. This AI for natural CH_4 working group aims to build a novel framework that integrates scientific knowledge and machine learning to harmonize simulated and observed datasets from global wetlands and soil sinks to quantify the spatial and temporal changes of global natural CH_4 fluxes. Specifically, we will harmonize every possible form of the global natural CH_4 datasets, including field-based CH_4 fluxes from chamber and eddy-covariance measurements and simulated CH_4 fluxes from bottom-up process-based models and top-down atmospheric assimilation models. As an output of this working group, we will generate and publicly share harmonized measurement datasets and global natural CH_4 flux products from 1980 to present.

Introduction and Goals

Atmospheric CH_4 grew at the fastest rate ever recorded in 2020-2022, reaching 1,920 parts per billion (ppb) in 2023[1]. CH_4 is a powerful greenhouse gas with a 100-year global warming potential 28-34 times that of CO_2 and could have caused nearly half of the global temperature increase since preindustrial times[2]. Slowing or reversing the accelerating growth in atmospheric CH_4 will require an improved understanding of the global CH_4 budget, which is currently underconstrained. Natural CH_4 budgets are responsible for ~40% of the total global CH_4 budgets but remain the most uncertain factor[3,4].

This AI for Natural Methane Working Group will improve our understanding of natural CH_4 budgets by harmonizing every possible form of datasets from global wetlands

and soil sinks. Wetland CH_4 emissions are the largest natural source of atmospheric CH_4, amounting to roughly 20–40 % (100-200 $TgCH_4yr^{-1}$) of global CH_4 budgets[5–7]. There are large disparities between a range of wetland CH_4 estimates from top-down and bottom-up models[3] due to large uncertainties in its spatial and temporal variability and overlooked processes. Natural CH_4 oxidation by microbes in upland soils is the second largest sink in the global CH_4 budgets, but its importance has been widely under-appreciated[8–10]. The current estimate of the global CH_4 soil sink is ~30 $TgCH_4yr^{-1}$ but with a huge uncertainty (7 to >100 $TgCH_4\ yr^{-1}$) from previous meta-analysis studies due to lack of measurements[11,12].

There are multiple forms of natural CH_4 datasets that can significantly reduce the uncertainties in our CH_4 estimation from wetlands and soil sink, including field-based CH_4 fluxes from chamber and eddy-covariance measurements, natural CH_4 fluxes simulated from top-down atmospheric inversions constrained by atmospheric observations, and simulated CH_4 fluxes from bottom-up process-based models. These datasets have different temporal and spatial scales, coverages, and implications on global CH_4 budgets. To synthesize these datasets, we propose to use a novel knowledge-guided machine learning (KGML) framework[13]. **We have a 2-year funded project from the Department of Energy (DOE) to develop a KGML framework for global CH_4 soil sinks (PI Oh, Senior Personnel Liu, McNicol, and Malone), which will leverage our working group activities.**

Proposed Activities

The objective of this working group is to synthesize multiple measured and simulated datasets using a KGML framework to better constrain natural CH_4 fluxes from wetlands and soil sinks (Fig. 1). This KGML framework will be designed to integrate scientific knowledge from bottom-up and top-down models, machine learning (ML) models and multi-source data through knowledge-guided architecture pretraining and training[14,15]. The growing field of KGML[16,17] provides a promising synthesizing method to combine multi-source datasets by adding knowledge-guided constraints into ML framework[17–21]. Specifically, we will harmonize the following four types of natural CH_4 datasets within the KGML framework.

First, simulated natural CH_4 fluxes from a bottom-up process-based model, the Terrestrial Ecosystem Model (TEM), will be used to pretrain the model[8,22]. The pretraining with modeled data is proven to help properly initialize the ML model and provide prior scientific knowledge in KGML[13,19,20,23,24]. Process-based models reflect our current scientific knowledge by explicitly applying equations and parameters from previous studies. We will use the bottom-up model estimates to pretrain the KGML model to incorporate the current scientific knowledge, and TEM is one of few biogeochemistry models that concurrently simulate wetland CH_4 emissions and soil CH_4 consumption[22,25] (**Dataset ID 1**).

Second, we will use simulated natural CH_4 fluxes from a global atmospheric CH_4 inversion model, the CarbonTracker-CH_4, to further pretrain the KGML[26,27]. Top-down atmospheric inversions constrain the surface emissions using atmospheric measurements of CH_4. The CarbonTracker-CH_4 further separates fossil, microbial, and fire emissions by assimilating atmospheric measurements of the stable carbon isotopic ratios of CH_4[28] and will incorporate the blended TROPOMI-GOSAT satellite retrievals[29]. To synthesize knowledge from atmospheric measurements of CH_4 and its isotopes, we will pretrain the KGML model using the simulated natural CH_4 fluxes from CarbonTracker-CH_4 (**Dataset ID 2**).

Third, we will incorporate measured net ecosystem CH_4 uptake fluxes collected from FLUXNET-CH_4[5,30]. FLUXNET-CH_4 provides a public global dataset of eddy covariance CH_4

flux measurements from 81 sites (~5 million fluxes) with very high temporal resolution (half-hourly or hourly), including 13 upland, 7 drained, 12 wet tundra, and 49 seasonally dry wetland sites. By synthesizing the FLUXNET-CH_4 dataset, we can provide high temporal information on natural CH_4 fluxes for KGML. **This synthesis will be the first to include a focus on drained or upland soils when net CH_4 uptake can be observed (Dataset ID 3).**

Lastly, natural CH_4 fluxes from chamber-based measurements from 1980s to present will be used to train the model. There are multiple chamber-based datasets available from previous literature. For example, Ni and Groffman (2018)[31] provide large datasets of CH_4 oxidation rates from previous chamber measurements of more than 300 soil sink studies. Also, multiple meta-analyses collected CH_4 fluxes from wetlands from more than 200 studies[32–34]. The collected chamber datasets of natural CH_4 fluxes will be synthesized to train the model to cover high spatial variability from 1980s to present (**Dataset ID 4-10**).

To ensure the readiness of the dataset for KGML development, a collaborative effort involving experts from both modeling and measurement fields will be made with in-depth discussions at the working group meetings. The discussions will focus on understanding the standards applicable to each dataset, including aspects such as format, definitions, units, resolution, handling of anomalies/missing values, and addressing uncertainties. Subsequently, the datasets will undergo harmonization, preprocessing, and storage processes, transforming them into Pytorch tensors aligned with a common standard conducive to rapid integration into ML models. Specifically, the most abundant temporal resolution within each dataset (e.g., monthly for chamber measurements and hourly for FLUXNET-CH4) will serve as the target harmonized resolution. Ultimately, data synthesis across varying temporal resolutions will be achieved through cross-scale KGML frameworks.

We will then explore various machine-learning methods, such as traditional recurrent, convolutional, and graph neural networks (RNN, CNN, GNN)[35], and advanced Gated Recurrent Unit (GRU)[36], Transformers[37], Graph Convolutional Neural Network (GCN), and Spatial Variability-Aware Neural Network (SVANN)[38] during the pretraining and training steps. We will reserve 30% of data from the collected FLUXNET-CH_4 measurements and chamber measurements to validate the estimated natural CH_4 fluxes. As an output of this working group, we will generate and publicly share harmonized measurement datasets and monthly global natural CH_4 flux products from KGML at 0.5-degree resolution from 1980 to present.

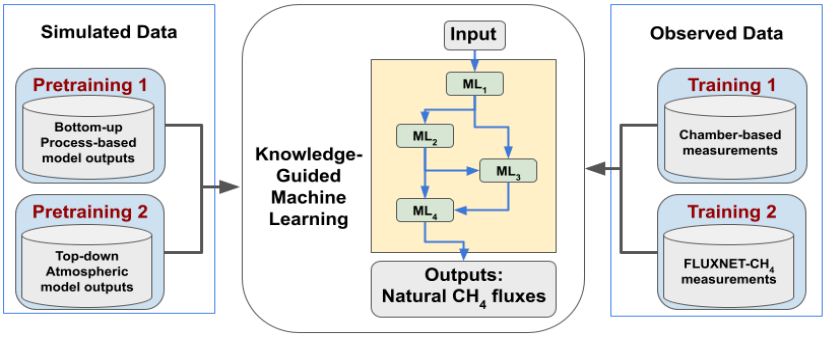

Figure 1. Overview of the KGML framework.

Advancing DEI

Our working group is committed to diverse representation, equitable outcomes, and an inclusive culture in scientific research. The PI and co-PIs have demonstrated DEI leadership in science communities by serving on several program and network committees,

including the Ameriflux Diversity, Equity, and Inclusion (DEI) networks, FLUXNET Early Career networks, and the multiple Mentoring programs[39]. Our working group leads are also passionate about advancing environmental data science via accessible education and training opportunities. Co-PI Malone has mentored students via the Environmental Data Initiative Fellowship program, while Co-PI McNicol has applied the principles of inclusive pedagogy to offer environmental data science courses at the University of Illinois Chicago.

In addition to ongoing contributions, we will design our working group to include diverse career stages, sectors, institution types, backgrounds and perspectives. Our group includes 2 graduate students, one of whom identifies as queer and Latino, and 10 early-career scientists, and all project leaders are early-career scientists. We are experts from diverse earth sciences, including atmospheric and flux measurements, ML, bottom-up and top-down modeling. When we have in-person and virtual meetings, our working group will design a Code of Conduct to foster a collaborative environment that respects and values diverse perspectives and identities, and where all members experience a sense of belonging and engagement.

Rationale for ESIIL support

Our working group's primary goal is to harmonize multiple datasets in order to gain a deeper understanding of the global CH_4 budget. This objective is in line with ESIIL's mission, which focuses on harnessing the abundance of environmental data and emerging analytical techniques to create science-based solutions for addressing critical challenges in biology and other environmental sciences. PI Oh is affiliated with the Cooperative Institute for Research in Environmental Sciences, while co-PI Sparkle Malone serves on the ESIIL Advisory Board. We are eager to support ESIIL's mission and goals, and we're also looking to partner with CyVerse to leverage their resources and customize their cutting-edge open science environment for our data synthesis initiatives.

Collaborations with other ESIIL activities

As an ESIIL working group, we will actively participate in various ESIIL's activities (Table 1). First, our working group will be involved in ESIIL's Hackathon, the Environmental MosAIc in 2024 and 2025. Graduate students and postdoctoral associates of our working group members will participate to join the Hackathon group, and other working group members will participate as a Hackthon mentor and expert panel. Second, our working group will actively participate in ESIIL's Innovation Summit in 2024 and 2025. We will form a global CH_4 breakout group and share diverse environmental challenges and opportunities. We will become part of the ESIIL Network through partnerships with individuals from various backgrounds who share research interests within diverse breakout teams and working groups.

Anticipated IT Needs

This proposed data synthesis will heavily rely on the development, pretraining, and training of KGML framework using harmonized datasets, which demands substantial high-performance computational resources. We will require access to a high-performance computing system equipped with multiple graphics processing units (GPUs). Around 3000 Nvidia A-100 GPU-hours will be needed for developing the KGML framework and implementing the KGML across scales and regions. Ensemble experiments will be also conducted in training phases. Also, given the large-scale synthesis of our datasets, storage capacity is paramount and the estimated storage need will be around 10 TB storage (refer to **Dataset table**). To meet these IT needs, we plan to leverage resources available through platforms like CyVerse (https://cyverse.org/), which provide specialized infrastructure and

support for data-intensive research projects. The resulting data and KGML model framework will be publicly available through open-data portals, such as DOE's ESS-DIVE or CyVerse.

Proposed Timetable

Our working group will be operating from February 2024 - January 2026 and will host 2 in-person meetings and 1-2 virtual meetings (Table 1). We will have our first in-person meeting at ESIIL on May 27-31, 2024. The primary objective of the first meeting is to pre-process and harmonize all the observed and simulated datasets of natural CH_4 fluxes. The working group will also share recent developments and knowledge gaps in global CH_4 budgets and work on a white paper or a commentary as an output of the in-person meeting. We will have our first virtual meeting on November 21-22, 2024, to share the results of the harmonized data product and develop the KGML framework. This virtual meeting will have breakout groups on wetlands and soil sinks separately to work on the detailed framework separately. Our second in-person meeting will be held at ESIIL on May 26-30, 2025. The primary objective of the second meeting is to share the preliminary results of the synthesized global natural CH_4 fluxes from KGML for suggestions. The working group will also share recent developments and knowledge gaps in global CH_4 budgets and will discuss ways to improve the harmonized datasets and KGML. If needed, our second virtual meeting will be held on November 20-21, 2025, to share our final datasets, white paper or commentary results, and working group activities.

Outcomes

As an output of this working group, we will generate and publicly share harmonized measurement datasets and monthly global natural CH_4 flux products from KGML at 0.5-degree resolution from 1980 to present (Table 1). The datasets and products will be shared on an open-data portal. Also, we will publish 1-2 white papers or commentaries discussing recent advancements and areas of knowledge gaps in global CH_4 budgets, as well as strategies to enhance our understanding through innovations in environmental data science.

Table 1. Proposed timeline of our working group activities.

	2024				**2025**			
Timeline	**May**	**Aug**	**Nov**	**Jan**	**May**	**Aug**	**Nov**	**Jan**
1st in-person meeting at ESIIL, Boulder								
Pre-Processing and Harmonizing Data								
1st Virtual meeting								
Developing KGML framework								
2nd in-person meeting at ESIIL, Boulder								
Estimating global natural CH_4 fluxes with KGML								
2nd Virtual meeting (if needed)								
Sharing our products of harmonizing datasets and KGML outputs and publishing white paper								
ESIIL Innovation Summit								
ESIIL Hackathon								

References

1. Stein, T. Greenhouse gases continued to increase rapidly in 2022. *NOAA Research News* (2023).
2. Masson-Delmotte, V., P. Z., A. Pirani, S., L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J B & R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R Yu And B Zhou. *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2021).
3. Saunois, M. *et al.* The Global Methane Budget 2000 – 2017. 1561–1623 (2020).
4. Oh, Y. *et al.* Improved global wetland carbon isotopic signatures support post-2006 microbial methane emission increase. *Communications Earth & Environment* **3**, 1–12 (2022).
5. McNicol, G., Fluet‐Chouinard, E. & Ouyang, Z. Upscaling Wetland Methane Emissions From the FLUXNET‐CH4 Eddy Covariance Network (UpCH4 v1. 0): Model Development, Network Assessment, and Budget Comparison. *AGU Advances*, 4(5), p.e2023AV000956.
6. Yuan, K. *et al.* Causality guided machine learning model on wetland CH_4 emissions across global wetlands. *Agric. For. Meteorol.* **324**, 109115 (2022).
7. Malone, S. L. *et al.* Gaps in network infrastructure limit our understanding of biogenic methane emissions for the United States. *Biogeosciences* **19**, 2507–2522 (2022).
8. Lee, J. *et al.* Soil organic carbon is a key determinant of CH_4 sink in global forest soils. *Nat. Commun.* **14**, 3110 (2023).
9. Murguia-Flores, F., Arndt, S., Ganesan, A. L., Murray-Tortarolo, G. & Hornibrook, E. R. C. Soil Methanotrophy Model (MeMo v1.0): A process-based model to quantify global uptake of atmospheric methane by soil. *Geoscientific Model Development* **11**, 2009–2032 (2018).
10. Oh, Y. *et al.* Reduced net methane emissions due to microbial methane oxidation in a warmer Arctic. *Nat. Clim. Chang.* **10**, 317–321 (2020).
11. Dutaur, L. & Verchot, L. V. A global inventory of the soil CH_4 sink. *Global Biogeochem. Cycles* **21**, 1–9 (2007).
12. Smith, K. A. *et al.* Oxidation of atmospheric methane in Northern European soils, comparison with other ecosystems, and uncertainties in the global terrestrial sink. *Glob. Chang. Biol.* **6**, 791–803 (2000).
13. Liu, L. *et al.* Estimating the Autotrophic and Heterotrophic Respiration in the US Crop Fields using Knowledge Guided Machine Learning. *ESSOAr* (2021) doi:10.1002/essoar.10509206.1.
14. ElGhawi, R. *et al.* Hybrid modeling of evapotranspiration: Inferring stomatal and aerodynamic resistances using combined physics-based and machine learning. *Earth and Space Science Open Archive* (2022) doi:10.1002/essoar.10512258.1.
15. Daw, A., Karpatne, A., Watkins, W., Read, J. & Kumar, V. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv [cs.LG]* (2017).
16. Karpatne, A., Kannan, R. & Kumar, V. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. (CRC Press, 2022).
17. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Computing Surveys* vol. 55 1–37 Preprint at https://doi.org/10.1145/3514228 (2023).

18. Irrgang, C. *et al.* Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence* **3**, 667–674 (2021).
19. Read, J. S. *et al.* Process‐guided deep learning predictions of lake water temperature. *Water Resour. Res.* **55**, 9173–9190 (2019).
20. Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sci.* **26**, 1579–1614 (2022).
21. Beucler, T. *et al.* Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. *Phys. Rev. Lett.* **126**, 098302 (2021).
22. Liu, L. *et al.* Uncertainty Quantification of Global Net Methane Emissions from Terrestrial Ecosystems Using a Mechanistically-based Biogeochemistry Model. *Journal of Geophysical Research: Biogeosciences* **125(6)**, e2019JG005428. (2020).
23. Khandelwal, A. *et al.* Physics guided machine learning methods for hydrology. *arXiv* (2020) doi:10.48550/arXiv.2012.02854.
24. Liu, L. *et al.* KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N_2O emission using data from mesocosm experiments. *Geoscientific Model Development* vol. 15 2839–2858 Preprint at https://doi.org/10.5194/gmd-15-2839-2022 (2022).
25. Zhuang, Q. *et al.* Response of global soil consumption of atmospheric methane to changes in atmospheric climate and nitrogen deposition. *Global Biogeochem. Cycles* **27**, 650–663 (2013).
26. Oh, Youmi, Lori Bruhwiler, Xin Lan, Sourish Basu, Kenneth Schuldt, Kirk Thoning, Sylvia E. Michel, et al. CarbonTracker CH4 2023. *NOAA Global Monitoring Laboratory* https://doi.org/10.25925/40jt-qd67 (2023) doi:10.25925/40jt-qd67.
27. Bruhwiler, L. *et al.* CarbonTracker-CH4: An assimilation system for estimating emissions of atmospheric methane. *Atmos. Chem. Phys.* **14**, 8269–8293 (2014).
28. Basu, S. *et al.* Estimating emissions of methane consistent with atmospheric measurements of methane and δ 13 C of methane. *Atmos. Chem. Phys.* **22**, 15351–15377 (2022).
29. Balasus, N. *et al.* A blended TROPOMI+ GOSAT satellite data product for atmospheric methane using machine learning to correct retrieval biases. *Atmospheric Measurement Techniques* **16**, 3787–3807 (2023).
30. Delwiche, K. B. *et al.* FLUXNET-CH 4: a global, multi-ecosystem dataset and analysis of methane seasonality from freshwater wetlands. *Earth system science data* **13**, 3607–3689 (2021).
31. Ni, X. & Groffman, P. M. Declines in methane uptake in forest soils. *Proceedings of the National Academy of Sciences* **115**, 8587–8590 (2018).
32. Kuhn, M. A. *et al.* BAWLD-CH 4: a comprehensive dataset of methane fluxes from boreal and arctic ecosystems. *Earth System Science Data* **13**, 5151–5189 (2021).
33. Wu, J., Cheng, X., Xing, W. & Liu, G. Soil-atmosphere exchange of CH4 in response to nitrogen addition in diverse upland and wetland ecosystems: A meta-analysis. *Soil Biol. Biochem.* **164**, 108467 (2022).
34. Hu, S. *et al.* Factors Influencing Gaseous Emissions in Constructed Wetlands: A Meta-Analysis and Systematic Review. *Int. J. Environ. Res. Public Health* **20**, (2023).
35. Zhang, S., Tong, H., Xu, J. & Maciejewski, R. Graph convolutional networks: a comprehensive review. *Computational Social Networks* **6**, 1–23 (2019).
36. Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv [cs.CL]* (2014).

37. Vaswani, A. *et al.* Attention is all you need. *arXiv [cs.CL]* (2017).
38. Gupta, J., Molnar, C., Xie, Y., Knight, J. & Shekhar, S. Spatial Variability Aware Deep Neural Networks (SVANN): A General Approach. *ACM Trans. Intell. Syst. Technol.* **12**, 1–21 (2021).

39. https://ameriflux.lbl.gov/community/group/diversity-equity-and-inclusion-committee/; https://ciresmentoring.colorado.edu/pages/about.php; https://fluxnet.org/community/ecn-early-career-scientist-network/.