DOE/SC-0213

AI4CH4

ARTIFICIAL INTELLIGENCE FOR THE METHANE CYCLE

2023 WORKSHOP REPORT

U.S. DEPARTMENT OF ENERGY | Office of Science

Biological and Environmental Research Program

March 2024

Artificial Intelligence for the Methane Cycle Workshop

March 3, 10, 17, 24, 2023

Convened by
U.S. Department of Energy
Office of Science, Biological and Environmental Research Program

Organizing Committee

Chair

Pamela Weisenhorn
Argonne National Laboratory

Co-Chairs

James Ang
Pacific Northwest National Laboratory

Jaydeep Bardhan
Pacific Northwest National Laboratory

Maxwell Grover
Argonne National Laboratory

Forrest M. Hoffman
Oak Ridge National Laboratory

Daniel Ricciuto
Oak Ridge National Laboratory

Charuleka Varadharajan
Lawrence Berkeley National Laboratory

Organizer

Olga Tweedy
U.S. Department of Energy, former AAAS Fellow

Point of Contact

Paul Bayer
U.S. Department of Energy

Additional Contributors

Dionysios Antonopoulos, Argonne National Laboratory; Gil Bohrer, U.S. Department of Energy; Christopher Henry, Argonne National Laboratory; Avni Malhotra, Pacific Northwest National Laboratory; Melanie Mayes, Oak Ridge National Laboratory; Gavin McNicol, University of Illinois–Chicago; Debjani Sihi, Emory University

About BER

The Biological and Environmental Research (BER) program supports transformative science and scientific user facilities examining complex biological, Earth, and environmental systems for clean energy and climate innovation. BER research seeks to understand the fundamental biological, biogeochemical, and physical principles needed to predict a continuum of processes occurring across scales, from molecules and genomes at the smallest scales to environmental and Earth system change at the largest scales. This research—conducted at universities, U.S. Department of Energy national laboratories, and research institutions across the country—is contributing to a future of reliable, resilient energy sources and evidence-based climate solutions.

This report is available at ess.science.energy.gov/ai4ch4

Suggested Citation

U.S. DOE. 2024. *Artificial Intelligence for the Methane Cycle*, DOE/SC-0213. U.S. Department of Energy Office of Science. https://doi.org/10.2172/2204972.

DOE/SC-0213

Artificial Intelligence for the Methane Cycle

Workshop Report

March 2024

U.S. DEPARTMENT OF ENERGY | Office of Science
Biological and Environmental Research Program

Contents

Foreword

This report derives from the March 2023 Artificial Intelligence for the Methane Cycle (AI4CH$_4$) virtual workshop, co-organized by staff from the Earth and Environmental Systems Sciences Division (EESSD), within the U.S. Department of Energy Biological and Environmental Research program (BER), and computational ecologist Dr. Pamela Weisenhorn from Argonne National Laboratory. AI4CH$_4$ provides a follow-up to the 2021 Artificial Intelligence for Earth System Predictability workshop series (ai4esp.org) co-organized by two DOE programs—BER and Advanced Scientific Computing Research (ASCR).

The purpose of AI4CH$_4$ was to more clearly demonstrate that artificial intelligence and machine learning (AI/ML) approaches could advance scientific understanding associated with one aspect of the global Earth system—the methane cycle. The workshop identified opportunities and challenges associated with better understanding the methane cycle as well as methane emission challenges across a range of spatial and temporal scales, from genomes to Earth system scales. AI4CH$_4$ also identified general AI/ML infrastructure challenges and social and cultural shifts that might be needed within methane science, data analytics, and modeling communities to fully realize the benefits of AI/ML approaches.

Executive Summary

Rapid advances in artificial intelligence (AI), machine learning (ML), and related advanced statistical approaches stand to accelerate progress toward scientific grand challenge goals by shifting the scientific research paradigm. The U.S. Department of Energy's (DOE) Biological and Environmental Research program (BER) is exploring opportunities and challenges in this emerging research area by sponsoring research and workshops, including the Artificial Intelligence for the Methane Cycle workshop ($AI4CH_4$) held over four virtual half days in March 2023.

AI is defined as any approach for building models from data to advance research objectives, alone or in conjunction with simulation, through techniques that enable computers to identify patterns, including ML, deep learning, and large language models. Developing and applying AI approaches in BER research can strengthen connections among datasets and individual steps of the integrated modeling and experimental (ModEx) framework (ess.science.energy.gov/modex), which decreases the time from scientific discovery to incorporation into predictive models. Furthermore, more rapid data assimilation and model development can provide greater focus in the design of future research studies and sampling campaigns leading to greater efficiency in creating novel insights.

Despite interest in the methane cycle from the Earth science community over the past decades, large uncertainties persist in global model estimates of land-atmosphere methane exchange. Methane is the second largest contributor to global warming, accounting for 20% of warming from greenhouse gasses and exhibiting a warming potential 27 to 30 times that of carbon dioxide over a 100-year time horizon. Model uncertainties reflect high variability of processes driving emissions, relative sparsity of process-relevant data, different measurement approaches and frequencies across the wide range of scales at which methane cycling is studied, and uncertainty in the measurements themselves. In addition, traditional numerical models insufficiently capture the multiscale nature of the methane cycle. A comprehensive solution demands the capability to find, link, leverage, and gap-fill existing datasets; incorporate these data into multiscale models; and evaluate the models against benchmark data.

The $AI4CH_4$ workshop focused on research needs and opportunities for applying AI in the specific context of the methane cycle. Workshop goals included identifying challenges and opportunities in data and modeling for the methane cycle and charting potential paths toward incorporating AI into future BER-supported research implementing the ModEx framework. Approximately 100 researchers from academia, industry, DOE, and other agencies contributed to the virtual workshop by submitting white papers or engaging in discussion sessions (see Appendix A: Agenda, p. 57; Appendix B: Participants, p. 58; and Appendix C: White Papers, p. 60).

Topical sessions focused on data and infrastructure needs for advancing predictive understanding of methane cycle components, with an emphasis on addressing the lack of closure between bottom-up and top-down methane emissions models, and the unique challenges posed by high spatiotemporal heterogeneity in individual methane cycling processes.

The $AI4CH_4$ workshop identified several key opportunities for AI application to methane cycle research:

1. **Enhance observation and experimentation.** In the field, sensor data can be assimilated into AI models, which are considerably faster and less computationally intensive than traditional models. This speed and energy efficiency enables real-time automated refinement of observation location and frequency based on sensor data and model output. Improved data collection leads to improved model resolution for processes with high spatiotemporal heterogeneity (e.g., methane ebullition). Similarly, autonomous

self-driving laboratories can potentially perform laboratory-based experimentation to evaluate AI model output. Output assimilated from prior experiments can be used to autonomously determine future experiments.

2. **Add contextual data to existing datasets.** Studies examining various aspects of the methane cycle have been conducted at many sites for decades. However, leveraging these existing datasets may require adding data or metadata (e.g., metagenomic data) that was not collected as part of the original study. AI-based data interpolation approaches, including generative adversarial networks, may be used to complete missing data. This approach increases the ability to perform data integration from disparate sources and improves data reusability, especially for purposes unanticipated at the time of sample, data, and metadata collection.

3. **Expand the findability and usability of data.** Application of AI can improve the exchange of data and model outputs, especially across scientific domains. Cross-domain exchange is necessary since methane cycle studies range from physiological investigations of microbial taxa to site-level flux measurements to global modeling efforts. However, the ability to leverage knowledge gained in one area to advance other efforts currently faces limitations. Large language models and retrieval-augmented generation can improve and automate ontology generation, and thereby improve the ability to find relevant data. The ability to use natural language queries can make data more findable and accessible to a broader range of users, including expert users outside the domain of the data generator.

4. **Optimize sampling strategies and experimental design.** The computational efficiency of AI methods enables rapid data assimilation which can be leveraged to help minimize parametric and structural uncertainty of models. It enables robust global sensitivity analysis, which can identify critical data needs and drive experimental design and sampling campaigns. Ultimately, these data can then be incorporated into local, regional, and global methane cycle models and thereby significantly improve model performance.

5. **Develop and support scientific workflows.** Progress toward AI and ML application in methane cycle research is slowed by computational, communication, ownership, and provenance challenges in developing, supporting, and maintaining scientific workflows that can integrate across the continuum from high-performance computing to cloud to edge compute capabilities. Scientific workflow progress could also be achieved by employing fully autonomous and self-driving sensing and experimental systems.

Application of AI and ML approaches across science domains within BER's Earth and Environmental Systems Sciences Division and Biological Systems Science Division offers exciting opportunities for accelerating progress towards BER's scientific grand challenge goals. Maximizing the application of AI and ML to gap-fill and link existing datasets can enable both experimentalists and modelers to more completely leverage existing knowledge about the methane cycle. Development and coupling of surrogate and hybrid AI models can improve the accuracy and efficiency of processes modeled over broad spatial and temporal scales, which remains challenging with traditional numerical modeling approaches. Particularly relevant to methane cycling and BER interests, AI and ML approaches could significantly advance the incorporation of individual microbial processes, such as methanogenesis and methanotrophy, into larger-scale models (BERAC 2017).

While workshop discussions examined challenges and opportunities in data and modeling for AI application in a methane cycle context, workshop participants also identified and discussed more general needs to support increased use of AI, including cultural shifts and computing infrastructure. The development of a community of practice and supporting infrastructure capabilities that enable increased application of AI to both data and modeling challenges promises benefits beyond methane cycling that could advance other BER-relevant scientific grand challenges.

Aerial view of a Spruce and Peatland Responses Under Changing Environments (SPRUCE) site. [Courtesy Oak Ridge National Laboratory]

1 | Introduction

The development of artificial intelligence (AI) has created new opportunities for advancing human understanding. AI modeling approaches possess an inherent ability to capture complex patterns in data spanning a vast spectrum of spatiotemporal measurement scales. Therefore, applying AI to diverse scientific research topics has the potential to rapidly improve predictions of Earth systems behavior. Such approaches can be especially effective for complex systems with numerous interacting components and nonlinear dynamics that are challenging to predict based on traditional mechanistic, process-based physical measurements and modeling.

This report, focusing on AI approaches in methane cycle research, defines AI as any approach for building models from data to advance research objectives, alone or in conjunction with simulation, using techniques (e.g., machine learning and deep learning) that enable computers to identify patterns following the scope set forth in DOE's AI for Science report (U.S. DOE 2020). Machine learning (ML) is defined as a subset of AI involving algorithms and statistical models where model performance can be improved over time through increased experience and data analysis.

AI and ML approaches are currently being developed and applied in methane cycle research. Although developments in generative AI, defined as techniques which can create text, images, data, and software by generalizing from patterns learned from large amounts of data (U.S. DOE 2020), have occurred rapidly, its applications in methane research are still limited. In general, biological applications of these approaches have advanced more rapidly than environmental and Earth science applications, in part due to the higher adoption of standardized data formats and conventions, such as those developed by the Genomic Standards Consortium (Yilmaz et al. 2011).

Artificial Intelligence Approaches

Classifier *(see Ch. 5)*

An algorithm that automatically orders or categorizes data into one or more groups or classes.

Synonym: classification model

Foundation Model *(see Ch. 1, 5, 6, 8)*

An AI model, often comprising trillions of parameters, that is trained on a broad range of data such that it can be applied across a wide range of use cases. Some foundation models can incorporate new information into their existing knowledgebase with minimal retraining, which is vital for adapting to the evolving nature of datasets in real-world scenarios.

Generative Adversarial Network (GAN)

(see Executive Summary and Ch. 2)

An unsupervised machine learning model in which two neural networks compete to generate increasingly authentic new data from a training dataset using deep learning methods.

Large Language Model (LLM)

(see Executive Summary and Ch. 1, 4, 6, 8)

A very large deep learning model, and a type of artificial neural network, pre-trained on vast amounts of data and notable for its ability to achieve general-purpose language outputs and generation.

Long Short-Term Memory Network (LSTM)

(see Ch. 1, 2, 4, 5)

A type of bi-directional artificial neural network, or recurrent neural network, which allows the output from some nodes to affect subsequent input to the same nodes as opposed to information flowing between layers in a forward-only direction. LSTM networks can reveal the importance of driving variables and their time dependencies in long time-series data.

Neural Network (NN) *(see Ch. 1, 2, 3, 5)*

A method in artificial intelligence that teaches computers to process data. Computers use NNs to learn from their mistakes and continuously improve accuracy of data analysis.

Examples include artificial neural networks, convolutional neural networks, and deep neural networks.

- ***Artificial Neural Network (ANN)***

  (see Ch. 1, 3, 5)

  A branch of machine learning used for solving artificial intelligence problems such as speech recognition, image analysis, and adaptive control. ANNs lie at the heart of deep learning algorithms.

  Synonyms: simulated neural networks, neural nets

- ***Convolutional Neural Network (CNN)***

  (see Ch. 5)

  Often utilized for classification and computer vision tasks, CNNs are distinguished from other NNs by their superior performance with image, speech, and audio signal inputs. CNNs provide a scalable approach to image classification and object recognition tasks.

- ***Deep Neural Network (DNN)*** *(see Ch. 5)*

  An ANN with multiple layers, each with a set of artificial neurons linked together, between the input and output layers. DNNs can theoretically map any input type to any output type, but they require many more examples of training data compared with other machine learning methods.

  Synonym: deep learning network

New data-driven AI approaches present challenges, including needing to develop explainable models to advance conceptual understanding and to characterize uncertainty in model predictions (see "Types of AI" sidebar, p. 3). While traditional modeling approaches attempt to capture all processes in a single model,

Types of AI

Anomaly Detection *(see Ch. 4, 8)*

The process of identifying data points, entities, or events that deviate from an expected range.

Synonyms: denoising, outlier detection

Causality-Guided Machine Learning

(see Ch. 2)

Using experimental and observational data to determine causal relationships between variables, then using this information to enable predictive machine learning models to provide insights into the relationships.

Synonyms: causal machine learning, knowledge-guided machine learning

Explainable Artificial Intelligence

(see Ch. 1)

A collection of tools and frameworks that enable understanding and interpretation of machine learning model predictions.

Ensemble Models *(see Ch. 3)*

A collection of multiple individual models to produce a final prediction. These individual models can be of the same type (homogeneous ensemble) or different types (heterogeneous ensemble). Ensemble methods are widely used in machine learning and artificial intelligence because they often improve predictive performance compared to using a single model.

Supervised vs Unsupervised Learning

(see Ch. 1)

Supervised learning is a machine learning model that trains on labeled datasets. These datasets "supervise" algorithms to classify data or predict outcomes. The labeled inputs and outputs enable the model to improve its accuracy over time. In contrast, unsupervised learning analyzes and clusters unlabeled datasets. Supervised models typically display higher accuracy but require work to label the data in advance.

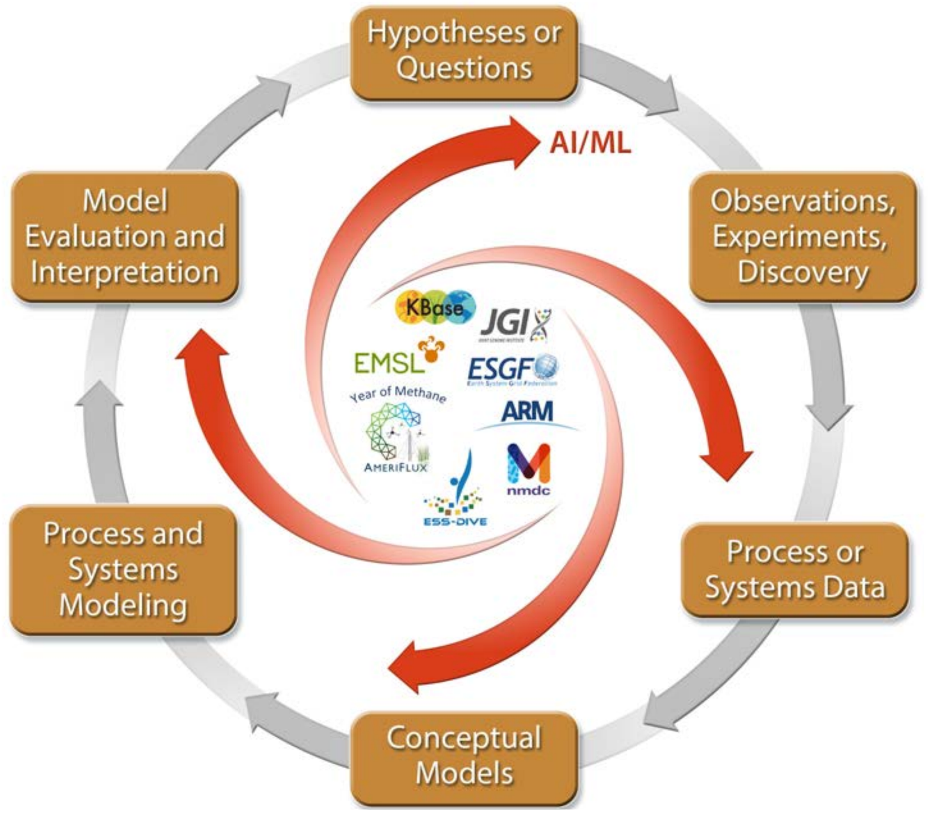
Surrogate Models *(see Ch. 1, 2, 4, 5, 6, 8)*

Simplified approximations of more complex, higher-order mathematical models that may have reduced accuracy but are computationally faster and cheaper to evaluate.

Synonyms: emulators, metamodels

hybrid models reduce computational cost by incorporating data-driven models for specific processes, often occurring at different temporal or spatial scales. These data-driven models may effectively predict system behavior even before a complete scientific understanding of the modeled process is developed. As with traditional process-based modeling approaches, AI and hybrid models can generate new scientific insights that provide a basis for additional laboratory and field experimentation and manipulation. Further consideration is needed to explore how to best leverage the strengths and address the limitations of AI and hybrid modeling approaches within an integrated modeling and experimental (ModEx) framework to advance research.

Research within the Biological and Environmental Research program's (BER) Environmental System Science program, for example, is guided by the ModEx framework (ess.science.energy.gov/modex). The ModEx framework integrates hypothesis-driven research (i.e., observations, experiments, and measurements) with modeling research that simulates the same processes. This integration supports incorporation of state-of-the-science research findings into modeling efforts, which in turn can be used to guide future research questions and directions. Developing and applying diverse AI approaches has the potential to impact each step in the ModEx cycle (see "Adapting the ModEx Framework to AI Models" sidebar, p. 4).



Adapting the ModEx Framework to AI Models

Artificial intelligence will accelerate work across all six stages of the model-experiment (ModEx) cycle:

- **Hypotheses or Questions:** (a) supplementing contextual metadata for existing datasets to leverage more data in developing hypotheses; (b) expanding data findability and usability. *(See Ch. 2, 7)*

- **Observations, Experiments, Discovery:** (a) pairing autonomous sensor systems with edge computing to quickly refine sampling location and frequency and improve process resolution (e.g., methane ebullition) with high spatiotemporal heterogeneity; (b) optimizing AI-driven sampling strategies and experimental design. *(See Ch. 3)*

- **Process or Systems Data:** (a) expanding data findability and usability; (b) ensuring quality assurance and control; (c) Improving data value through classification, outlier detection, and clustering, such as by integrating imaging data (e.g., computed tomography and neutron tomography) with geochemical flux data. *(See Ch. 4)*

- **Conceptual Models:** (a) connecting ontologies from different domains; (b) identifying rough contours of new interactions. *(See Ch. 2)*

- **Process and Systems Modeling:** (a) advancing Earth system models through surrogate and hybrid modeling approaches, especially with difficult-to-integrate measurements across scales; (b) parameterizing and reducing uncertainties in bottom-up process models. *(See Ch.5)*

- **Model Evaluation and Interpretation:** (a) assimilating data, analyzing sensitivity, and quantifying uncertainty, such as with surrogate models and hybrid surrogate/physics models; (b) identifying/characterizing model–measurement deviations. *(See Ch. 6)*

A strengthened connection between ModEx and data collection, management, and use requires reimagining the data lifecycle and how the scientific community captures and shares data. This is increasingly important considering the need to generate large and high-quality datasets for AI model development, such as foundation models and large language models.

Exploring AI for Energy and Environmental Science

Recognizing the potential for a paradigm shift in scientific discovery and prediction, DOE held a 17-session virtual workshop series, Artificial Intelligence for Earth System Predictability (AI4ESP), from October to December 2021 (U.S. DOE 2022). The workshop was jointly sponsored by BER's Earth and Environmental Systems Sciences Division (EESSD) and the Advanced Scientific Computing Research (ASCR) program within the DOE Office of Science. The workshop's goal was to pursue a shared agenda of applying advanced statistical approaches and AI to rapidly improve Earth systems predictability at spatial scales ranging from microbes through ecosystems to the globe, and at temporal scales spanning minutes to centuries. AI approaches produce models with improved abilities to resolve complex and nonlinear systems, meeting an increasing need to quickly make and refine high-resolution predictions that provide actionable information. Topics discussed throughout the workshop included (1) capturing heterogeneity in relevant variables and processes, (2) overcoming the difficulty associated with observing and predicting extreme events, (3) managing and analyzing immense volumes of data across a variety of ecosystems, and (4) launching a major effort to identify robust, interdisciplinary scientific approaches that integrate human activities (U.S. DOE 2022).

The AI4ESP workshop highlighted needs for:

- Large, curated datasets for model training.
- Incorporating AI to enhance observations.
- New mathematical approaches tailored to sparse data and extreme events.
- Novel approaches that address interpretability and potential physical inconsistencies of traditional ML model outcomes, driving the need for hybrid models.
- Innovative and consistent approaches to representing model uncertainties and trustworthiness.
- Software infrastructure to support hybrid model components across major Earth and environmental system science codes.
- Efficient and interoperable frameworks and architectures that provide access to data and model resources across organizations (U.S. DOE 2022).

Specifically, the workshop recognized a need for a supporting framework to lower community-wide barriers to access and bridge domain-specific needs for data generation, standards, synthesis, and model development.

Another workshop, Artificial Intelligence and Machine Learning for Bioenergy Research (U.S. DOE 2023) was held jointly by BER's Biological Systems Science Division (BSSD) and the DOE Bioenergy Technologies Office (BETO) in August 2022. This workshop focused on bioenergy research and using data-driven approaches in parallel with traditional hypothesis-driven approaches to accelerate design and optimization of biological systems and processes for biotechnology innovation. The workshop spanned BSSD-funded research from enzymes to interacting biological systems and particularly assessed the potential to advance biological understanding and engineering capabilities through development of AI/ML-driven autonomous laboratory systems. It highlighted biosystems design requirements for development of data and computational infrastructure.

Three critical needs and opportunities were identified:

- Availability of robust, high-quality data with complete metadata.
- Development of improved or novel algorithms to meet the specific needs of the biotechnology community.
- Further development and use of laboratory automation approaches.

Methane: A Crosscutting Research Opportunity

Building off the momentum of previous workshops, a follow-on workshop, Artificial Intelligence for the Methane Cycle (AI4CH$_4$), was held over four virtual half days in March 2023. The workshop, detailed in this report, focused on unique needs and opportunities arising from using AI models to advance understanding of the methane cycle at scales from genomes to the global Earth system.

The extent to which research questions discussed in the workshop intersect with BER priorities is well-captured by the BER Advisory Committee's grand challenge in microbial to Earth system pathways: "Define the levels of biological organization most relevant to scaling from single cells to ecosystems and global cycles; capture how that organization varies in time and space; and identify critical interactions that dictate the rates of carbon, nutrient, and energy transformation in different environments" (BERAC 2017). In addition, methane research touches on several grand challenges outlined in the EESSD Strategic Plan for 2018 to 2023 (U.S. DOE 2018), including:

- **Biogeochemistry.** Advance a robust, predictive understanding of coupled biogeochemical processes and cycles across spatial and temporal scales by investigating natural and anthropogenic interactions and feedbacks and their associated uncertainties within Earth and environmental systems.
- **Drivers and Responses in the Earth System.** Advance next-generation understanding of Earth system drivers and their effects on the integrated Earth-energy-human system.
- **Data–Model Integration.** Develop a broad range of interconnected infrastructure capabilities and tools that support the integration and management of models, experiments, and observations across a hierarchy of scales and complexity to address EESSD scientific grand challenges.

One established research thrust within BER's funded programs is to develop a predictive understanding of the global carbon cycle. The cycle consists of carbon stocks and fluxes that interact across the atmosphere, biosphere, lithosphere, and ocean, impacting both biological systems and climate. Application of AI-based approaches can accelerate predictive understanding of the carbon cycle by improving identification of complex patterns in sparse datasets and reducing computational challenges associated with modeling of organisms and processes across markedly different spatial and temporal scales. As an important component of the carbon cycle, the methane cycle is an ideal case study to explore the opportunities and challenges in applying AI because its biology is fairly well-understood and because of its outsized impact on climate feedbacks.

> *As an important component of the carbon cycle, the methane cycle is an ideal case study to explore the opportunities and challenges in applying AI.*

Natural methane emissions are strongly linked to specific groups of microorganisms, a subset of which are well-studied, and methane release is further regulated by microbe–microbe and microbe–plant interactions. Many physicochemical conditions, including oxidation-reduction (redox) potential, carbon substrate availability, and alternative electron acceptor availability, influence both methane production and consumption. Ebullition and the physical conditions that favor it can influence methane release from flooded soils and sediments to the atmosphere due to methane's low water solubility.

Once released into the atmosphere, methane is a greenhouse gas with 27 to 30 times the radiative forcing of carbon dioxide on a 100-year time horizon (IPCC 2021), resulting in larger-scale feedbacks to the climate system. Methane cycling has been of intense interest to the Earth science community. Despite this interest and body of knowledge, large uncertainties persist in global model estimates of land-atmosphere methane exchange (Saunois et al. 2020). These uncertainties are partly due to the high spatial and temporal variability of methane emissions (Bousquet et al. 2006; Rosentreter et al. 2021). This variability has many causes:

- Physiologic potential and environmental sensitivity of microorganisms.
- Significance of microbe–microbe and plant–microbe interactions to methane production, consumption, transport, and release.
- Importance of abiotic processes, including pore structure, in mediating methane release.
- Ecological disturbances or perturbations.

These uncertainties in land-atmosphere methane exchange are exacerbated by the relative sparsity of process-relevant environmental data and metadata, differences in measurement approaches and frequencies across the wide range of scales at which methane cycling is studied, and uncertainty in the measurements themselves. The biological and environmental processes underlying high methane flux variability are difficult to incorporate into traditional numerical models. This is due, in part, to limited observations and environmental context for measurements on which to base such models, as well as variation in the spatial and temporal scales at which processes occur (Lan et al. 2021; Bridgham et al. 2013).

Workshop Overview

Approximately 100 scientists with various expertise from 30 national laboratories, universities, industry, and other federal agencies contributed to the workshop, either through a pre-workshop call for white papers or direct participation. The first day consisted of plenary presentations, initial brainstorming, and ideation. The remaining days were divided into topical sessions focused on (1) predictions from fundamental microbiology, (2) environmental controls and empirical relationships, (3) field measurements and observations, (4) data–model integration challenges, (5) multiscale modeling, and (6) key conclusions. Participants discussed knowledge gaps and identified key scientific questions suitable for advancement using AI, characteristics and challenges of relevant data and models, opportunities for development of related algorithms and infrastructure, and data products and needs currently limiting progress in the field. Specifically, attendees discussed the importance of:

- Understanding the physiology, activity, and impact of methanogens and methanotrophs at ecosystem scales and larger.
- Understanding how plant traits affect soil methane release.
- Understanding how pore structure and pore-scale processes affect methane efflux.
- Connecting measurements and insights between laboratory and field.
- Capturing hot spots and hot moments in methane fluxes.
- Integrating field methane measurements across scales, from chambers to towers to satellites.
- Establishing benchmarks for model development and intercomparisons.

An NGEE-Arctic researcher takes measurements amid a swarm of mosquitoes. [Courtesy Lawrence Berkeley National Laboratory]

2 | State of the Science

Methane is a critical component of the carbon cycle with significant implications for Earth's climate. It is the second largest contributor to global warming, accounting for 20% of warming from greenhouse gasses (Ciais et al. 2013; Kirschke et al. 2013). On a per-ton basis, methane's global warming potential, a measure of a greenhouse gas' capacity to absorb thermal radiation, is 27 to 30 times that of carbon dioxide (CO_2) over a 100-year time horizon (IPCC 2021). Because methane is short-lived in the atmosphere (i.e., 12-year half-life), its impact is even more pronounced at shorter time scales, exhibiting 80 to 83 times the warming potential of CO_2 over a 20-year time horizon (IPCC 2021). These characteristics make methane a reasonable and effective short-term target for reaching climate goals where CO_2 targets are currently falling short. Reducing methane emissions to 40% to 70% of 2020 levels by 2030 is critical to staying below the Paris Agreement limit of 1.5°C global average surface warming above pre-industrial temperatures (IPCC 2021; UNEP 2023; Rogelj and Lamboll 2024). Rapid advancement in both conceptual and predictive understanding of the methane cycle is required to reach this goal.

Mitigating methane emissions requires an ability to differentiate between human-driven and natural emissions, as well as to predict emissions changes due to changing environmental conditions. Recent methane levels are more than 150% pre-industrial levels, with 2020 and 2021 setting records for largest annual increases in methane concentrations (15.3 parts per billion and 17 ppb, respectively) since direct air measurements began in 1983 (NOAA 2022). Annual changes in atmospheric methane concentration have varied greatly since the early 1990s (see Fig. 2.1, p. 10). An unexpected increase in methane concentrations began in 2007 following a stable period from 1992 to 2007. Large uncertainties in both measured and modeled methane emissions and sinks have

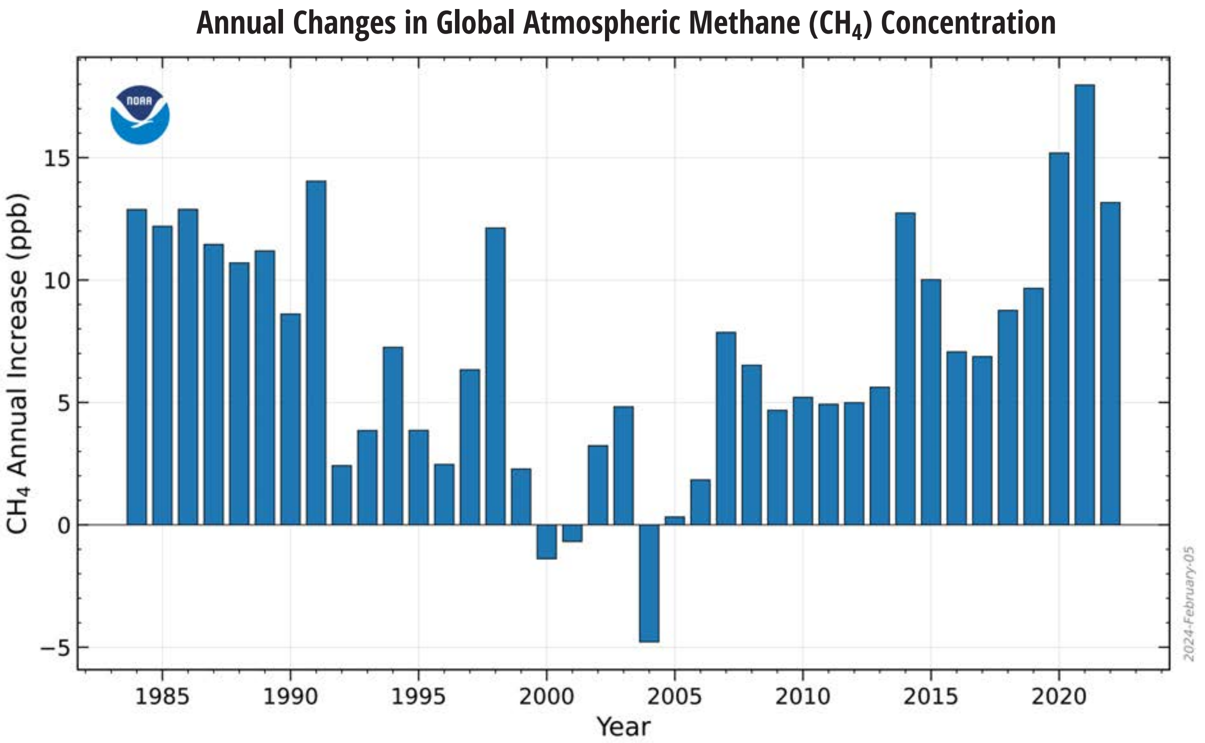


Fig. 2.1. Annual Changes in Atmospheric Methane (CH_4) Based on Globally Averaged Marine Surface Data. This graph captures three distinct periods of global changes in atmospheric methane concentration. First, a period of generally decelerating annual increases from the mid-1980s to 1992. Second, a period of relative stability between 1992 and 2007 where methane concentrations both increased and decreased at lower levels. Finally, a period of accelerating increases in methane concentration since 2007. [Courtesy NOAA. gml.noaa.gov/ccgg/trends_ch4]

limited the ability to develop causal links to explain this recent trend of accelerating increases in methane concentration.

Current and Projected Methane Sources

Methane sources include wetlands, freshwater systems (e.g., lakes, rivers, streams), wildfires, oil and gas infrastructure, landfills, cattle, and agriculture (see Fig. 2.2, p. 11). With continued climate change, thawing permafrost and dissociating methane hydrates from marine sediments may greatly increase methane emissions, though this remains uncertain (Ketzer et al. 2020; Malakhova and Golubeva 2022; Schuur et al. 2022) and would not be expected to occur in this century (IPCC 2021).

Wetlands and inland lakes are dominant natural sources of global methane emissions, with roughly a third of natural emissions coming from each system (Saunois et al. 2020; Rosentreter et al. 2021; IPCC 2021). However, estimates of the contributions and relative importance of these methane sources remain uncertain, partly due to the dynamic and complex nature of these systems and the high levels of uncertainty in capturing the areal extent of wetlands and lakes (Pham-Duc et al. 2017; Zhang et al. 2021) and partly due to challenges in scaling from ground measurements to atmospheric observations (Saunois et al. 2020).

Despite pronounced wetland loss due to human activities, the effects of climate change on temperature and precipitation regimes are driving increasingly high overall wetland methane emissions (Peng et al. 2022;

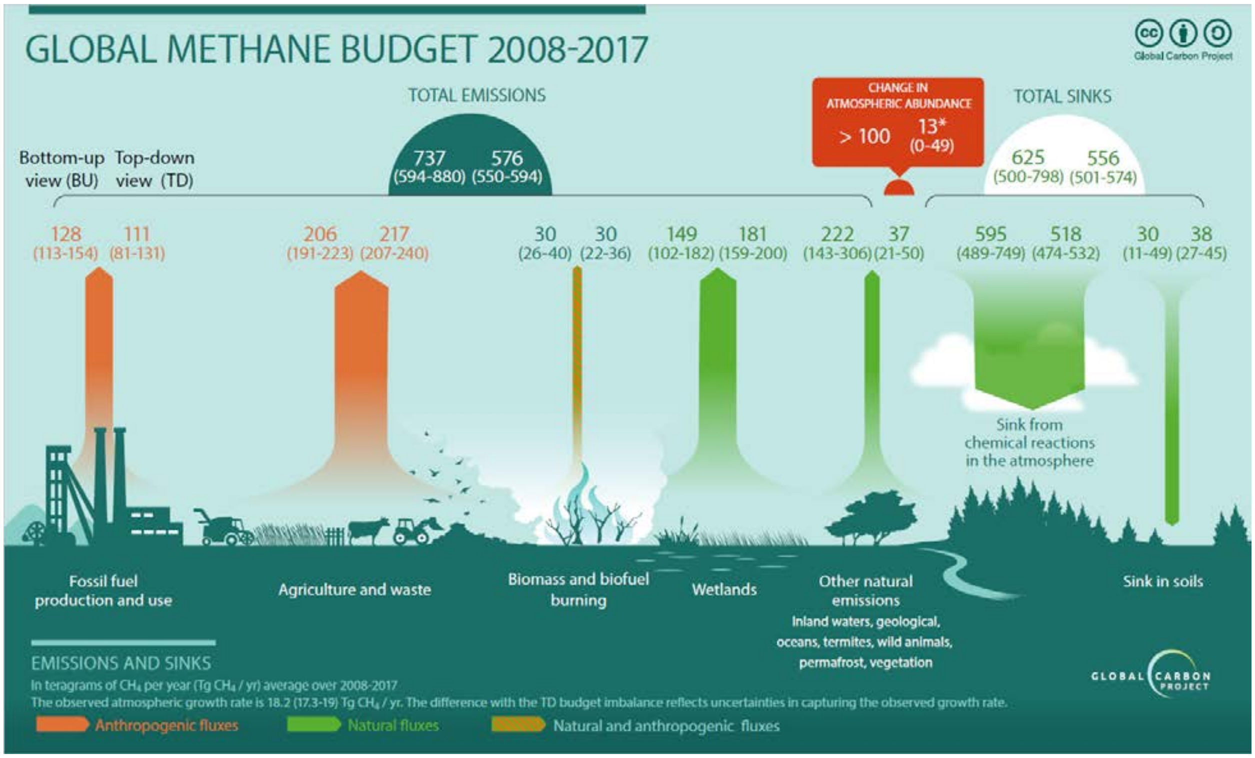

GLOBAL METHANE BUDGET 2008-2017
Global Carbon Project
TOTAL EMISSIONS
Bottom-up view (BU)
Top-down view (TD)
737 (594-880)
576 (550-594)
CHANGE IN ATMOSPHERIC ABUNDANCE
> 100
13* (0-49)
TOTAL SINKS
625 (500-798)
556 (501-574)
128 (113-154)
111 (81-131)
206 (191-223)
217 (207-240)
30 (26-40)
30 (22-36)
149 (102-182)
181 (159-200)
222 (143-306)
37 (21-50)
595 (489-749)
518 (474-532)
30 (11-49)
38 (27-45)
Sink from chemical reactions in the atmosphere
Fossil fuel production and use
Agriculture and waste
Biomass and biofuel burning
Wetlands
Other natural emissions
Inland waters, geological, oceans, termites, wild animals, permafrost, vegetation
Sink in soils
EMISSIONS AND SINKS
In teragrams of CH_4 per year (Tg CH_4 / yr) average over 2008-2017
The observed atmospheric growth rate is 18.2 (17.3-19) Tg CH_4 / yr. The difference with the TD budget imbalance reflects uncertainties in capturing the observed growth rate.
Anthropogenic fluxes
Natural fluxes
Natural and anthropogenic fluxes
GLOBAL CARBON PROJECT

Fig. 2.2. Discrepancies Between Bottom-Up and Top-Down Models of Global Methane. Flux estimates from the Global Carbon Project's Methane Budget for the period 2008 to 2017 capture the large variation and discrepancy between bottom-up and top-down modeling approaches. For example, bottom-up models estimate global methane emissions rates at 737 teragrams/year, but top-down models estimate 576 teragrams/year. In this figure, the pair of numbers accompanying each arrow represents bottom-up (left) and top-down (right) estimates of methane emissions or sinks, with the range for each average shown in parentheses. Primary emissions include fossil fuel production and use, agriculture and waste, biomass and biofuel burning, wetlands, and other natural sources. Sinks include soils and atmospheric chemical reactions. [Reprinted under a Creative Commons License from Global Carbon Project]

Fluet-Chouinard et al. 2023; Rößger et al. 2022). Researchers broadly agree that methane emissions from wetlands will increase through the 21st century (IPCC 2021), although the magnitude of this change remains highly uncertain (Koffi et al. 2020; Chang et al. 2023). Recent model projections accounting for methane-climate feedbacks suggest that increased wetland emissions in response to altered temperature and precipitation regimes could nullify an estimated 25% to 40% of targeted reductions in anthropogenic emissions needed to meet the temperature goals of the Paris Agreement (Zhang et al. 2023), thereby necessitating even larger reductions in anthropogenic emissions.

Continued warming is also expected to increase methane emissions from lakes, especially those located in northern latitudes or experiencing eutrophication (Beaulieu et al. 2019; Jansen et al. 2022; Zhuang et al. 2023). Emissions from inland waters comprise the largest source of uncertainty in the methane budget, ranging from 6 to 185 teragrams (Tg) CH_4/yr over the past 20 years (Johnson et al. 2022). These estimates are impacted by high spatial and temporal variation in fluxes, uncertainty in lake areal extent, and a relatively small number of observations (IPCC 2021). Critically, current top-down budgets do not account for inland waters, contributing to the large gap in bottom-up and top-down estimates of "other" sources in the Intergovernmental Panel on Climate Change's (IPCC) methane budget (Table 5.2 in IPCC 2021).

Another significant natural methane source is wildfires, which are predicted to increase in frequency regionally with climate change (Turner et al. 2019; Pausas and Keeley 2021). Forest management practices, such as fuel reduction, can influence the total amount and proportional release of methane during wildfires (Volkova et al. 2014). While wildfires directly contribute pyrogenic methane, they also impact soil carbon pools and can indirectly affect the spatial pattern and extent of methane release or uptake for years post-fire (e.g., Davidson et al. 2019; Wilkinson et al. 2023).

Dominant sources of anthropogenic methane emissions include cattle farming and fossil fuel infrastructure (see Fig. 2.2, p. 11; Saunois et al. 2020; IPCC 2021). These substantial sources are expected to respond to climate change differently than natural emissions. For example, emissions from oil and gas production are complex and poorly constrained, varying as a function of atmospheric conditions and state of infrastructure (e.g., well construction, age, and maintenance; see Krofcheck and Nole white paper, p. 87). Reducing emissions from certain anthropogenic sources (e.g., methane leaks from oil and gas infrastructure) can be accomplished using proven technologies, but identifying and responding to leaks first requires extensive monitoring and data processing infrastructure (UNEP 2023). Meanwhile, efforts to measure and curb emissions from livestock and agriculture, which are the largest anthropogenic contributors to methane emissions (see Fig. 2.2, p. 11), have been more limited, with a less clear path forward (Reisinger et al. 2021).

Current and Projected Methane Sinks

Relative to methane sources, non-atmospheric methane sinks have received much less research attention. The primary methane sink is atmospheric degradation by hydroxyl radicals (see Fig 2.2, p. 11; Kirschke et al. 2013). The second largest sink is aerobic and anaerobic methane oxidation by methanotrophic bacteria in soil and aquatic ecosystems, though this microbiological sink is understudied and potentially underestimated (Zhao et al. 2019; Jing et al. 2020; Feng et al. 2023). Uptake of methane by soil microbes accounts for 30 to 38 Tg/yr, or roughly 5% of total methane sinks (see Fig. 2.2, p. 11).

Soil methanotroph abundance varies on the global scale with climate (particularly mean annual temperature and mean annual precipitation), soil properties (i.e., pH and total organic carbon), and vegetation cover (Ding 2024; Heděnec et al. 2024). More research is needed to understand how localized soil sinks respond to altered precipitation, temperature, and atmospheric methane concentrations. Some models and meta-analyses show increases (e.g., Gatica et al. 2020; Murguia-Flores et al. 2021) while others show decreases (e.g., Ni and Groffman 2018). Notably, pan-Arctic models found that the activity of high-affinity methanotrophic bacteria can potentially double the predicted soil sink strength in this region (Oh et al. 2016; Oh et al. 2020).

Soil methane sinks exist across many ecosystems, but cropland soil sinks require special consideration as changes are not only influenced by the atmospheric and environmental context but also by dynamic management strategies (Kim et al. 2021). Accounting for the management component of the soil sink requires understanding human behavior in response to a shifting economic context. Management strategies can potentially increase the cropland soil sink (e.g., Runkle et al. 2019; Kim et al. 2021). Conversely, increased soil moisture resulting from increased frequency of intense rainfall events could shift cropland systems to methane sources (Cowan et al. 2020; see Morris et al. white paper, p. 89).

While there is broad consensus that methane release from natural systems will increase over the next decade, there is no such consensus on the magnitude or direction of change in the global methane sink.

Predictive Framework

The role of methane as a greenhouse gas has long been recognized, and many numerical models have been developed over the past 40 years to capture environmental controls of methane fluxes from terrestrial systems (Xu et al. 2016). Recent microbial- to ecosystem-scale models of methane emissions have demonstrated the importance of incorporating methanogenic substrate production and availability, microbial population size and activity, plant-mediated transport of methane, and methane ebullition to improve predictive accuracy (Song et al. 2020; Ricciuto et al. 2021; Sihi et al. 2021). Ocean-atmosphere modeling components incorporating microbial processes have also been developed (Reinhard et al. 2020). Incorporating microbial drivers into models has improved predictability across scales, from ecosystem to land surface models.

Within natural systems, developing a predictive understanding of the methane cycle is hindered by its complex nature. The many biotic and abiotic drivers and their non-linear interactions contribute to high spatial heterogeneity in localized methane emissions (Sturtevant et al. 2016; see Yuan et al. white paper, p. 64). For example, soil pore structure can influence soil moisture content, thereby influencing microbial habitat, redox dynamics, and both methane production and consumption (see Mayes et al. white paper, p. 97). Changes in microbial community composition in soils can influence landscape-scale methane emissions, but predictive understanding of the interactions between these biotic and abiotic components is lacking (He et al. 2015; Hartman et al. 2017).

Data Availability

At the global scale, many international efforts to increase methane monitoring are underway, including the United Nations Environment Programme's International Methane Emissions Observatory (IMEO). Satellite and aerial measurement campaigns that were recently launched (e.g., Sentinel-5 and Carbon Mapper) or will soon launch (e.g., GOSAT-GW and MethaneSAT) will increase the accuracy, spatial resolution (i.e., down to 1 km^2), and temporal frequency of atmospheric methane measurements. Additionally, programs like IMEO and Carbon Mapper aim to provide open access to near real-time quantitative methane emission data. Alongside these improvements in global atmospheric methane data measurements, improvements are underway regarding measurements of surface characteristics (e.g., land temperature and vegetation) from satellites (e.g., PlanetScope and Hydrosat) and new data products (e.g., Moon 2022).

The high spatial and temporal variability in global to local methane emissions underscores the need for diverse data collected from many high-quality, spatially representative sites (IPCC 2019; see Yuan et al. white paper, p. 64). The AmeriFlux network launched its "Year of Methane" campaign in 2019 to build support for collecting methane flux data from a more diverse range of sites. This effort resulted in the release of the FLUXNET-CH4 dataset which includes methane flux data from 79 global sites, with high representation of freshwater wetlands (Delwiche et al. 2021). FLUXNET-CH4 has played a critical role in advancing predictive understanding of the methane cycle and is referenced throughout this report. Recently, McNicol et al. (2023) developed a random-forest-based upscaling model using the FLUXNET-CH4 dataset to provide bottom-up estimates of methane flux. The model succeeded in capturing patterns in extratropical methane fluxes, but more long-term methane flux monitoring is needed to capture tropical ecosystems (Delwiche et al. 2021; McNicol et al. 2023; Yuan et al. 2023).

Recent modeling work has demonstrated the importance of including data on microbes and plant traits in predictions of methane emissions. Efforts to capture paired microbial and biogeochemical data are underway, such as in the high latitudes (e.g., Barret et al. 2022), but more biological data and models are needed to capture the critical effects of microbes and plants on methane cycling (see Bueno de Mesquita and Tringe white paper, p. 70). A methane-focused effort comparable to the Genome Resolved Open Watersheds database release in DOE's Systems Biology Knowledgebase (Borton et al. 2022) can potentially provide insight into genome–phenotype–environment relationships. Further, this presents a key opportunity to apply advanced AI approaches to integrate genomic data with ecosystem-level measurements (see Xu and Rodrigues white paper, p. 80).

Scientific Challenges

Substantial existing challenges in integrating and translating data across scales are recognized and discussed throughout this report. Examples of these challenges include the shifting relevance of individual factors across scales and the difficulties in incorporating processes occurring at vastly different spatial and temporal scales into a single modeling framework. These challenges are not unique to the methane cycle and underlie many BER grand challenges related to "improving the predictive power of Earth system models" and understanding the influence of microbial communities on soil and plant systems and their subsequent effects on regional and global environments (BERAC 2017).

Within the context of the methane cycle, workshop attendees identified and discussed four scientific challenges: (1) developing an integrated understanding of methane-cycle biology, (2) identifying interspecific and abiotic–biotic effects on plant and microbial function, (3) modeling connections between microscale processes and larger-scale process rates, and (4) resolving discrepancies in global wetland methane emission estimates between bottom-up and top-down models.

Scientific Challenge 1: Developing an Integrated Understanding of Methane-Cycle Biology

Natural methane emissions result from a balance between methane production, consumption, transport, and release to the atmosphere. Methanogens and methanotrophs directly contribute to the methane cycle but their activity can be strongly affected by interactions with other microbial taxa (e.g., Megonigal et al. 2004; Nozhevnikova et al. 2020; Metcalfe et al. 2021; Vigderovich et al. 2023). Plants can also influence methanogens, methanotrophs, and other microbial community members through litter and root exudates. Further, plant adaptations to flooding (i.e., highlighting the importance of aerenchyma) can influence methane transport from the soil and affect the spatial distribution of oxidation-reduction conditions (i.e., redox potential) within the soil through radial oxygen loss. Thus, understanding the biotic component of the methane cycle requires physiological examination of both plants and microbes, how their activity is impacted by interspecific and interkingdom interactions, and their response to shifting biotic and abiotic conditions.

1.1: Acquiring a physiological understanding of microbial populations and the mechanisms controlling their functioning

While many factors influence the high spatiotemporal variability of methane flux measurements, microbes and their interactions must be recognized for their critical importance in methane production and consumption. Indeed, microbial methanogenesis could not occur without interactions between methanogenic taxa and other community members (Megonigal et al. 2004). Methane consumption that occurs near sites of active methanogenesis can substantially reduce methane release and is often dependent on interspecific interactions. For example, methane removal from ocean sediments occurs primarily through the activity of anaerobic methanotrophic archaea that must directly transfer electrons to a syntrophic partner bacterial species to support their metabolism (Skennerton et al. 2017).

> *The important role of microbiota and their interactions is more pronounced in the methane cycle due to the relatively few taxa that directly participate compared to other biogeochemical cycles.*

The important role of microbiota and their interactions is more pronounced in the methane cycle due to the relatively few taxa that directly participate compared to other biogeochemical cycles. This leads to an increased importance of individual microbial taxa, each with a unique set of physiological tolerances, interspecific interactions, and metabolic dependencies. Ecosystem and land-surface modeling results indicate that the accuracy of methane predictions improves with inclusion of microbe-specific parameters (Song et al. 2020; Ricciuto et al. 2021; Sihi et al. 2021; Yuan et al. 2021).

One of the greatest impediments to integrating microbial mechanisms into climate and ecosystem models is the continued inability to translate environmental genomic data into complete, accurate, and predictive mechanistic models of microbial taxa and communities. Yet several emerging approaches can enable a more complete understanding of the activity of methanogens, methanotrophs, and other key microbial community members under a range of natural and experimental conditions.

One challenge is overcoming the many errors and knowledge gaps that persist in functional genome annotation (Warren et al. 2010; Dimonaco et al. 2022). Application of deep learning methods, in particular, may help address this challenge by opening new avenues for analyzing the function of unknown genome sequences by providing insights into three-dimensional protein shapes, which are crucial for understanding their functions and interactions. For example, the AI program AlphaFold accurately predicts protein structures, enabling researchers to decipher the roles of proteins with unknown genome sequences and advance biological understanding (Jumper et al. 2021; Varadi et al. 2022).

AI application and related experimental improvements can help address the challenge of functional genome annotation by delivering (1) new automated laboratory and rapid phenotyping methods that greatly improve the quantity of experimental data used to drive validation and correction of gene annotations and microbial phenotype predictions (see Ch. 3: Observations, Experiments, and Discovery, p. 19; U.S. DOE 2023) and (2) AI methods that improve the integration and consistent accurate propagation of functional annotations across reference genomes. Further, AI methods can be combined with metabolic modeling approaches to fill knowledge gaps that can easily break the predictive power of a metabolic model operating on its own. They can also provide greater insight into the physiology of single key microbial community members, including methanogens and methanotrophs (Kavvas 2020; Bi et al. 2023).

While novel culturing approaches are progressing (e.g., Wu et al. 2020; de Raad et al. 2022), most microbial species remain unculturable, leaving only metagenomic data as a tool to reconstruct microbial genomes from the environment. This problem is especially significant in the study of the methane cycle, which involves diverse and unknown methanogenic and methanotrophic taxa (e.g., Narrowe et al. 2019; Smith and Wrighton 2019; Bay et al. 2021; Ellenbogen et al. 2023). Challenges in metagenome binning and assembly of complex microbiomes typically result in mostly low-quality, highly incomplete genomes. This lack of high-quality genomes complicates metabolic modeling of methane cycling communities and slows understanding of energy flow and conservation in these communities. In-depth understanding of culturable microbial communities is possible but requires substantial investment of time and resources (e.g., Orphan et al. 2022).

1.2: Acquiring a physiological understanding of plant populations and the mechanisms controlling their functioning and interactions with microbes

Methane production and consumption are microbial processes, but plants and their traits play important roles in transport and regulating net methane emission, both locally and globally. On plot to site scales, plant traits have long been recognized as important predictors of methane emission across ecosystems (e.g., Whiting and Chanton 1993; Malhotra and Roulet 2015; Knox et al. 2021), but their impact is less clear in fine-scale methane processes along the plant-soil-microbe interface. Plants provide substrate for methanogenesis in the form of dead plant material (litter) for decomposition and live plant products such as root exudates (Lai 2009; Sutton-Grier and Megonigal 2011; Bridgham et al. 2013). Plant tissues (e.g., aerenchyma) also serve as conduits for methane transport from soils to the atmosphere and atmospheric oxygen to soils, the latter leading to radial oxygen loss and potentially causing increased methane oxidation in anaerobic settings (van Bodegom et al.

2001; Noyce et al. 2023). It remains unclear how methane-relevant plant traits differ among plant species and to what extent this trait variation is genetically determined (Takahashi et al. 2013; Yoo et al. 2015; Wany and Gupta 2018).

It further remains unclear how plants influence methane processes differently between ecosystem types and under changing environmental conditions. For example, global warming is expected to increase above- and below-ground plant productivity in some wetland ecosystems (Noyce et al. 2019; Hanson et al. 2020; Malhotra et al. 2020). The effect on net methane emissions could be positive due to increased substrate availability and plant transport or negative due to increased rhizosphere oxygenation (Hopple et al. 2020; Noyce et al. 2023). Examples from a boreal peatland, a tropical peatland, and a salt marsh provide some insights into plant trait–methane linkages, but not enough measurements exist across ecosystems to parameterize plant–trait linkages with methane production in local- and global-scale models (Xu et al. 2016; Yuan et al. 2023).

While recent and soon-to-launch satellites will improve the capture of methane and satellite-observable plant traits (e.g., productivity), many plant traits critical to the methane cycle (e.g., rooting traits and aerenchyma) cannot be observed in this way. Increased measurements and trait distribution models are needed to better capture these traits.

Scientific Challenge 2: Identifying Interspecific and Abiotic-Biotic Effects on Plant and Microbial Function

Organisms behave differently in natural environments than in laboratory settings, and many challenges exist in translating laboratory-derived scientific understanding into field settings. Thus, while it is important to develop a mechanistic understanding of what organisms can do and the conditions under which they grow or perform specific functions (e.g., methanogenesis, methanotrophy), it is also important to understand their distribution and how environmental context, including soil pore structure and moisture, influence growth and activity. Microbial activities involving methane production and consumption are highly influenced by surrounding soil moisture content and pore structure. Spatiotemporal heterogeneity in soil biophysical conditions at the microsite scale regulates hot spots and hot moments of methanogenic and methanotrophic activities and ultimately, net methane emissions at the ecosystem scale (Sihi et al. 2021; Lacroix et al. 2023). Hence, imaging and modeling tools that predict microbial responses to variant soil conditions may aid in predicting gross and net methane fluxes.

Understanding the complex relationships among environmental conditions, interspecific interactions, and plant and microbe physiological tolerances is critical to predicting their contribution to methane cycling. Plant species distribution and niche modeling comprise an established area of research that can be leveraged to understand these complex relationships for plant communities. However, such information is limited for microbiomes. Given the complexity of these relationships, direct generation of meaningful and testable hypotheses under field conditions is difficult. One path forward is through developing robust neural network models using genomic content as predictors of microbial community composition in response to environmental controls (e.g., Larsen et al. 2011; Mallick et al. 2019; Reiman et al. 2021). Evaluating these models, built from high-throughput laboratory data and tested on field data, is critical to determining the transferability of lab-derived knowledge into more complex environments.

Distributed shotgun metagenomic sequencing efforts, such as the Genome Resolved Open Watersheds database (Borton et al. 2023), could provide initial data for evaluating such models, although iterative laboratory research and validation for specific functional guilds and ecosystem types may be required. Notably, because the presence or abundance of microbial taxa may serve as predictor variables, such large metagenomic datasets may also help identify genetic markers and microbes that impact methane emissions but happen to fall outside of existing focal guilds (Khan et al. 2023), providing new avenues of exploration for laboratory studies.

A similar approach could link plant and microbial community composition to methane cycling. However, few available field datasets capture both methane flux and plant and microbial species composition (see Bueno de Mesquita and Tringe white paper, p. 70). This presents a major data gap (see Ch. 4: Data Curation, Integration, and Products, p. 25), and recent modeling efforts have underscored the need for plant and microbial trait data. To address this challenge, generative adversarial networks (GAN; see "Artificial Intelligence Approaches" sidebar, p. 2) could leverage existing field data on microbial community composition across a range of environmental conditions. GANs would use this data to generate training datasets of simulated microbial community composition based on environmental conditions for sites that have methane flux data but lack microbial data. However, considerable additional data are likely needed both for robust evaluation of these models themselves and any downstream models using these data.

A further challenge remains, however, in that most sites with long-term, ecosystem-scale methane flux measurements (e.g., FLUXNET-CH4) capture only net flux rates. These data reflect the balance between methane production and consumption processes but obscure linkages between the underlying biology and measured rates, hindering modeling efforts. To address this, data-constrained Bayesian (e.g., Ueyama et al. 2022; Ueyama et al. 2023) or causality-guided machine learning (e.g., Yuan et al. 2022; see Zhu et al. white paper, p. 62; see "Types of AI" sidebar, p. 3) models could further partition net methane flux rates into production, consumption, and transport processes.

Scientific Challenge 3: Modeling Connections Between Microscale Processes and Larger-Scale Process Rates

Understanding the influence of microbial processes at larger, aggregate scales is critical to unraveling complex methane cycle dynamics. Achieving this goal requires bridging the gap between microscale processes and larger-scale ecosystem behavior. A scientific challenge remains in identifying the key elements of soil pores and microbial communities that are crucial for generating methane cycle predictions at larger scales and defining the level of detail required to accomplish this goal.

Incorporating metagenomic information related to processes like methanogenesis and methanotrophy into large-scale numerical models poses a significant challenge. The challenge results, in part, from the orders of magnitude difference in scales between microbial and global processes and the difficulties in reducing large metagenomic datasets into smaller, better targeted parameter sets. Recent approaches have examined microbial functional groups (e.g., Song et al. 2020; Ricciuto et al. 2021; Sihi et al. 2021), but this remains an active area of study (see Song et al. white paper, p. 99; Xu and Rodrigues white paper, p. 80). One approach is to apply neural networks, including long short-term memory networks (see "Artificial Intelligence Approaches" sidebar, p. 2), to develop surrogate models (see Oh et al. white paper, p. 74). Such neural network models offer both computational efficiency and scalability.

Hot spots and hot moments in methane production, transport, and release further complicate traditional modeling approaches. These localized emissions bursts play a pivotal role in methane dynamics, making it vital to address them accurately. However, attempts to model them, including by employing AI approaches, is hindered by data scarcity. More advanced and flexible imaging and modeling techniques are needed to handle the dynamic nature of soil pores and improve sampling design and sensors (see Ch. 3: Observations, Experiments, and Discovery, p. 19).

Addressing the broader challenge of modeling across scales, applying model consolidation or ensemble approaches (see "Types of AI" sidebar, p. 3) may be useful to address parametric and structural uncertainty (see Ch. 5: Multiscale Modeling, p. 33). The combination and evaluation of various models results in a better predictive outcome and a more comprehensive understanding of ecosystem dynamics, especially in the context of microbial contributions to large-scale phenomena.

Scientific Challenge 4: Resolving Discrepancies in Global Wetland Methane Emission Estimates Between Bottom-Up and Top-Down Models

Large discrepancies in predicted global methane fluxes persist between bottom-up and top-down models. Bottom-up methane models estimate methane emissions by aggregating ground-based data from individual sources or activities and then use inventories, scaling parameters, and statistical approaches to aggregate them to larger scales. Meanwhile, top-down models are based on atmospheric measurements of methane and use inversions of atmospheric transport models to identify sources of methane flux. Bottom-up models estimate global emissions rates at 737 (range: 594 to 880) Tg CH_4/year, but top-down models estimate 576 (550 to 594) Tg CH_4/year. Bottom-up models estimate sink rates at 625 (500 to 798) Tg CH_4/year, but top-down models estimate 556 (501 to 574) Tg CH_4/year (see Fig. 2.2, p. 11).

Workshop discussions and white paper topics revealed that much of this discrepancy may result from incomplete bottom-up models. One challenge with bottom-up model development is data sparsity (e.g., soil carbon sinks; see Oh et al. white paper, p. 74) and biased distribution of available data (see Delwiche et al. white paper, p. 93). In the case of biased data, models can be developed to compensate for the bias, but additional data are needed to validate these models.

Another challenge is the inability to access or leverage existing data, either due to challenges in cross-domain, interdisciplinary science (e.g., plant traits and microbial activity; see Scientific Challenge 1, p. 1, and Scientific Challenge 2, p. 9) or issues with proprietary data (e.g., see agroecosystems in Morris et al. white paper, p. 89; see oil and gas infrastructure in Krofcheck and Nole white paper, p. 87).

A further challenge is the ability to capture the high degree of spatiotemporal heterogeneity in environmental conditions that influence the methane cycle, including flooding (see Stachelek et al. white paper, p. 72) or wildfires (see Li et al. white paper, p. 85).

A final challenge is partitioning existing emissions data into processes that better align with the conceptual understanding captured in bottom-up models, such as between natural and anthropogenic emissions data (see Krofcheck and Nole white paper, p. 87) and between individual process components of net methane flux (see Zhu et al. white paper, p. 62). Many of these gaps are progressing toward resolution, albeit slowly.

Researchers measure methane and carbon dioxide emitted from coastal forest soil. [Courtesy Pacific Northwest National Laboratory]

3 | Observations, Experiments, and Discovery

Current measurement-based estimates of global methane flux contain large uncertainties due to a lack of sufficient measurements across scales, including local, regional, and global scales. Field- and laboratory-based approaches to methane data collection include point-source measurements that are spatially localized and often offer poor temporal coverage. Examples include soil incubations from field samples, bubble traps, and flux chambers. Micrometeorological flux towers (see Fig. 3.1, p. 20; Morin et al. 2017) offer high temporal resolution but low spatial resolution. Other measurements provide improved spatial coverage but only for snapshots in time, including vehicle-mounted sensors (e.g., Picarro, Aeris, and Licor) and drone or aircraft-based sampling. Remote sensing offers large regional spatial coverage, but its limitations include poor temporal coverage, poor spatial resolution, and high detection limits. Poor resolution complicates identification and quantification of confounding methane sources, such as landfills, livestock, and oil and gas refineries.

Datasets from these measurements are critical to artificial intelligence and machine learning (AI/ML) workflows because they serve as benchmarks for potential algorithms aimed at solving methane-specific problems. AI-enabled measurement platforms, including edge computing and high-throughput, autonomous laboratories, can help address some of the limitations with measurement collection. Edge computing can focus measurements, enabling better capture of critical data including hot spots and hot moments, and autonomous laboratories

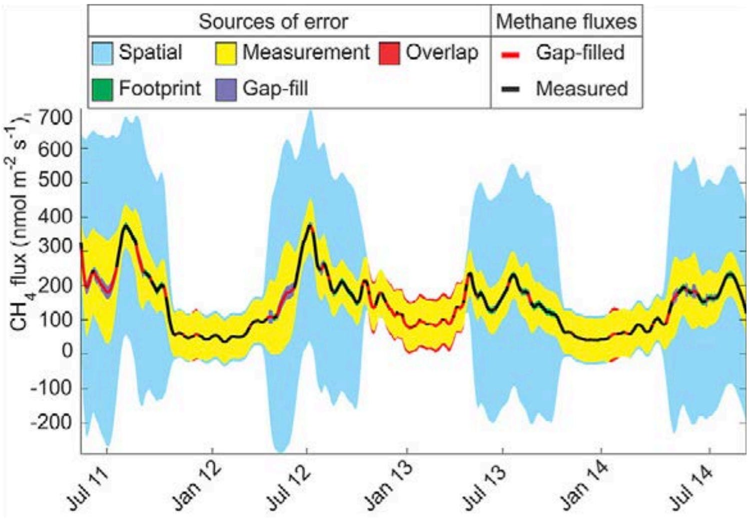


Fig. 3.1. Evaluating Sources of Error in Gap-Filled Data Using Artificial Intelligence (AI). Large uncertainties in methane emissions rates (yellow and cyan shading) make it difficult to determine whether carbon dynamics in an urban freshwater wetland offset net greenhouse warming effects. Two different methane flux measurement techniques, eddy covariance and chamber measurements, were used in tandem to correct for their respective limitations and achieve a better estimate of net methane flux (black line). An AI model using an artificial neural network (ANN) was used to gap-fill the methane flux observations (red line), and Monte Carlo simulations of the AI ANN model were used to determine the uncertainty of these estimates due to observation errors (yellow shading) and spatial heterogeneity of fluxes and sampling locations (cyan shading). [Morin, T. H., et al. 2017. "Combining Eddy-Covariance and Chamber Measurements to Determine the Methane Budget from a Small, Heterogeneous Urban Floodplain Wetland Park," *Agricultural and Forest Meteorology* **237-238**, 160-170.]

can increase data generation from limited field samples, providing better biological insights into methane processes.

Observations Needed Across Scales

Workshop participants identified several gaps in multiscale, multidisciplinary data needed to accurately inventory top-down and bottom-up methane emissions, advance modeling and analysis, and improve predictive understanding of the methane cycle. Many of the data collection technologies needed to fill these gaps could benefit from extensive network coverage in remote areas through advanced telemetry and wide-area networking (e.g., satellite internet service or 5G cellular service). Filling these gaps would accelerate progress toward predictive understanding of the methane cycle:

- **Improved Measurements of Methane Hot Spots and Hot Moments.** Methane release is highly heterogeneous in space and time, and hot spots and hot moments make substantial contributions to overall emissions. Flexible, high-throughput, and automated methane measurements are required across a comprehensive set of biological and environmental conditions. AI-based instrumentation, such as unmanned aerial vehicles (i.e., UAVs, drones) that follow inverse modeling measurements, may be particularly useful in capturing this variability. New, on-the-ground robotic technologies combined with autonomous sensing platforms can be applied as well.
- **Higher-Resolution Spatiotemporal Methane Measurements.** Data gaps often exist in methane flux measurements from eddy covariance towers and chambers. Filling these gaps is crucial to enabling scaling from intra-daily temporal scales to seasonal and annual flux estimates. AI approaches, including artificial neural networks (ANN), can be used to perform temporal gap-filling of continuous flux measurements (e.g., eddy covariance measurements; see Fig. 3.1, this page). Spatial coverage of methane measurements can be gap-filled as well, expanding flux measurements across environmental gradients, including belowground properties, to better capture a range of spatial variability.
- **Spatially and Depth-Resolved Data on Subsurface Properties.** Methane production,

consumption, transfer, and release are influenced by soil properties, such as soil pore structure, mineral chemistry, and organic matter chemistry. New approaches or sensors are needed to obtain spatially resolved data that are difficult to observe with remote sensing, such as soil and subsurface properties. For example, measurements of microsite heterogeneity in drivers of methane production and oxidation are limited but needed to capture spatial heterogeneity in models and explore spatial distribution of taxa. While additional manual measurements and sampling campaigns, such as the Molecular Observation Network (MONet, www.emsl.pnnl.gov/monet), are needed, ensemble modeling approaches can be leveraged to generate spatially resolved data products (e.g., Mishra et al. 2021).

- **Microbial Community Information Paired with Rate and Function Measurements.** Microbes are key players in the production and consumption of methane, but more data and better understanding of microbial processes across a range of sites and conditions are needed to understand how their physiology and community interactions impact methane emissions. High-throughput automated experiments can be leveraged to better understand microbial physiology and species interactions, while advanced techniques like isotope pool dilutions and gas push-pull can be used to quantify gross rates of production and oxidation in field settings with more complex communities. Field validation of findings from these high-throughput automated experiments at diverse, long-term field sites will factor crucially into knowledge transfer from laboratory to field and larger-scale models.

Sampling Location and Design

Observational strategies should be targeted across spatiotemporal scales, from microsite-level methane heterogeneity within local ecosystems to the global methane budget. Methane measurements focused on local-scale features are critical for identifying and understanding local sources and sinks. In addition to measurements across scales, robust model–data intercomparisons require well-defined surface flux benchmark datasets. Such datasets can enable seamless comparisons across scales, connecting local ecosystems to the global Earth system.

Methane measurements across every ecosystem would be ideal, but they are unfeasible. Thus, a strategy is required to determine which areas are most critical for measurement and observation. This approach includes using a combination of bias assessments in sample distribution and quantified measurement and modeling uncertainties to predictively identify high-value locations and times for additional measurements. Additional remote sensing instrumentation, such as satellite-based or airborne platforms, is also needed. Blending these remote sensing and local *in situ* observations is critical to filling spatial and temporal data gaps and may require focused sampling campaigns.

For smaller-scale observations, from single sites to ecosystems, challenges arise when determining where and when to collect data. One potential target area for new flux tower measurements is along hydrological gradients, which are sources of many model uncertainties (see Delwiche et al. white paper, p. 93). Another critical measurement location is the transition between soil sinks and sources, specifically redox dynamic zones where soil fluctuates between oxic and anoxic conditions. Understanding microsite distribution and response to changing moisture and redox conditions is critical to predicting hot spots and hot moments of methane production and consumption. Integrating automated sensing of real-time methane emissions at the ecosystem-scale and soil biophysical conditions with laboratory-based snapshots of soil pore structure and microsite distributions would enable capture of the spatial and temporal dynamics of underlying methane cycle processes.

For larger-scale observations, a key issue is global methane budget uncertainty. Large-scale measurement strategies require satellite remote sensing platforms, such as the TROPOspheric Monitoring Instrument (TROPOMI), which can collect high-resolution measurements around the globe with regular frequency. The richness of these remote sensing datasets provides an opportunity to use explainable AI to examine soil, thermal, hydrological, and geochemical processes contributing to methane dynamics. ML can also be used to blend local ecosystem measurements with remote sensing observations in a reproducible and physics-constrained manner, enabling improved methane flux retrieval.

Focused sampling strategies and data representation are needed to improve capacity for blending remote sensing and *in situ* local measurements, which enables seamless ingest, prediction, and evaluation of various AI/ML algorithms aimed at bridging the gap between these measurements.

Improved Data Collection Technologies

In addition to key regions of opportunity, new measurement platforms are also available. These include UAVs, edge sensing and computational platforms, and autonomous laboratories. UAVs enable high spatiotemporal sensing for scanning entire local systems and could collect methane fluxes both near and above Earth's surface. While the new platforms hold great promise, an important challenge to applying them is the remoteness of many high-methane-producing regions, which complicates data transfer and ML model application in the field.

Edge Computing for Methane Measurements

Edge computing platforms offer development and application of increasingly robust methane measurement techniques (see Fig. 3.2, p. 23). These systems enable transparent end-to-end workflows, including consistent quality control methods and dynamic products based on environmental features. Edge computing serves as a means of controlling the sensing and actuation of data collection, which enables environmental observations at a level of detail appropriate to existing conditions. For example, edge computing can increase data resolution (i.e., increase sampling frequency) during low atmospheric pressure or storm events when methane ebullition is more likely to occur. Edge computing can be leveraged by combining high-density forward-looking infrared (FLIR) cameras with image processing to detect methane hotspots and adjust sampling location accordingly. This approach achieves computational efficiency and meets data reduction goals by processing only the data that are necessary and relaying only model outputs to primary data storage systems.

Edge computing can also be applied beyond measurement collection. Measurements can be blended with other data sources, such as FLUXNET-CH4 or existing U.S. Department of Agriculture crop-type databases, to execute pre-trained AI models or test recently developed algorithms. Pattern recognition could be used as well, relaying discovered methane signatures to the edge computing system and triggering process models to run when necessary. One potential model based on these signatures is the Massively Parallel Reactive Flow and Transport Model for Describing Subsurface Processes (PFLOTRAN; Hammond et al. 2014). Data from such computationally expensive models can be intelligently compressed and transferred to larger cloud- or high-performance computing data centers using 5G technology, where available (see Mudunuru et al. white paper, p. 101).

Autonomous Laboratories for Model Training and Evaluation

A laboratory analogue to AI-driven or enhanced sensor platforms is the autonomous laboratory. Autonomous laboratory systems are envisioned to combine the power of high-throughput data generation with AI-enabled analysis and automated experimental execution. The opportunity to establish such approaches to large-scale data generation from complex systems makes them attractive in the large combinatorial spaces typically found in biology, such as examining pairwise and multi-species interactions in communities with hundreds to thousands of species.

Empirical data generation from biological processes in the laboratory can be time-consuming and may not scale with the myriad factors influencing methane generation and consumption. Autonomous laboratories can help explore concentration gradients and combinations of biotic, abiotic, and interspecific factors influencing organism growth or process rates involved in methane cycling. However, examining the methane

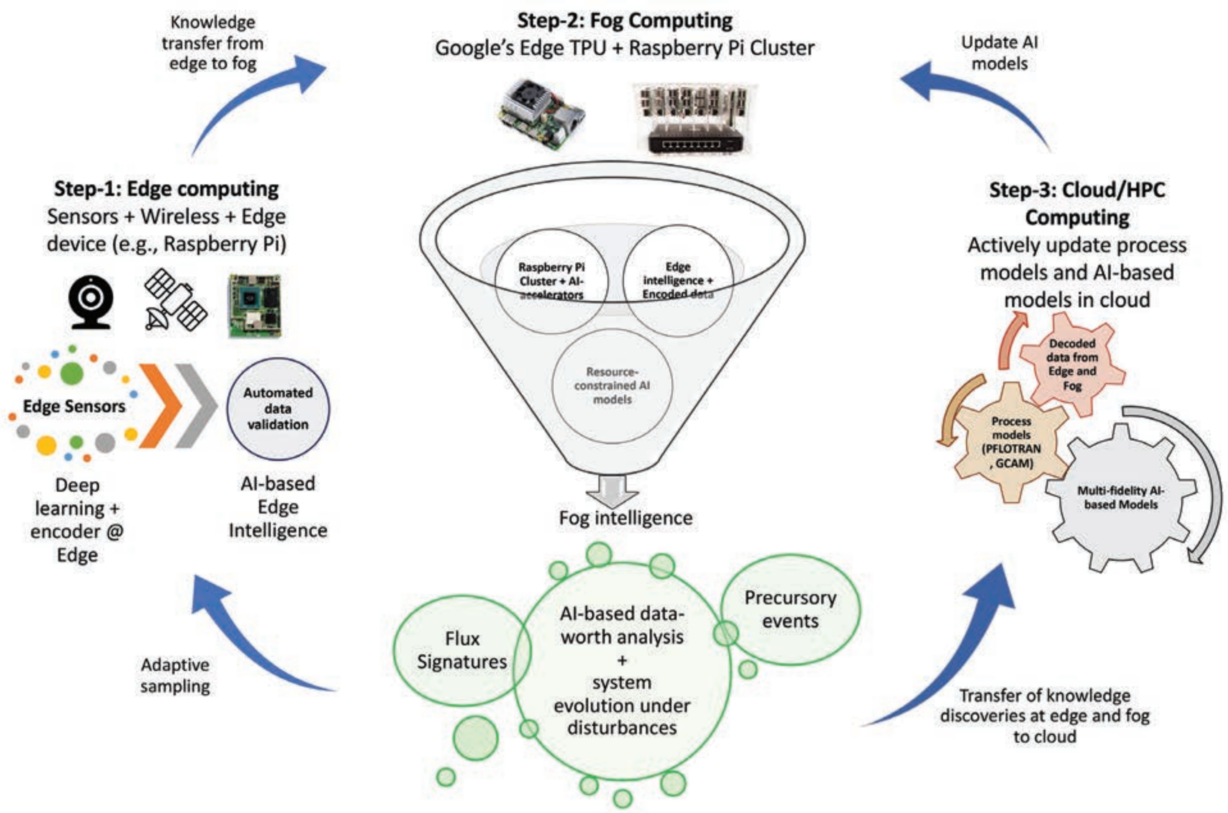

Knowledge transfer from edge to fog
Step-2: Fog Computing
Google's Edge TPU + Raspberry Pi Cluster
Update AI models
Step-1: Edge computing
Sensors + Wireless + Edge device (e.g., Raspberry Pi)
Raspberry Pi Cluster + AI-accelerators
Edge intelligence + Encoded data
Resource-constrained AI models
Step-3: Cloud/HPC Computing
Actively update process models and AI-based models in cloud
Decoded data from Edge and Fog
Process models (PFLOTRAN , GCAM)
Multi-fidelity AI-based Models
Edge Sensors
Automated data validation
Deep learning + encoder @ Edge
AI-based Edge Intelligence
Fog intelligence
Flux Signatures
AI-based data-worth analysis + system evolution under disturbances
Precursory events
Adaptive sampling
Transfer of knowledge discoveries at edge and fog to cloud

Fig. 3.2. Edge Computing Defined. This illustration highlights what is meant by "edge computing," or the practice of bringing computation closer to data collection. Typical non-edge-computing paradigms involve collecting observations (often in remote locations), transferring data to a large computing center, then performing scientific analyses. In edge computing paradigms, sensors (see Step 1) are connected directly to a tensor processing unit (TPU)/central processing unit (CPU), which can be used to deploy and train complex AI/ML workflows. A TPU/CPU established near the edge avoids a data resampling or reduction step which is often needed to efficiently transfer data to cloud and high-performance-computing (HPC) centers. Instead, edge computing is capable of handling full data resolution or, in cases where HPC is necessary, it could aid in advanced AI/ML data reduction techniques. Both capabilities would benefit the analysis of methane observations. [Mudunuru, M. K., et al. 2021. *EdgeAI: How to Use AI to Collect Reliable and Relevant Watershed Data, AI4ESP-1095*. U.S. Department of Energy Office of Science, Biological and Environmental Research Program.]

cycle via autonomous laboratory approaches poses additional challenges, both because methanogens require low-to-no oxygen environments and because methane itself exists in a gaseous state as a substrate or end-product at ambient temperatures. These challenges may be better addressed in miniaturized contexts (e.g., microfluidic-inspired reactors) where better control of local experimental conditions can be implemented. Advanced imaging techniques also may assist with phase change observations. Additionally, workflow automation is not solely physical but also encompasses information and data processing. Autonomous laboratory systems can serve as ideal platforms to train and evaluate models (U.S. DOE 2023).

Two key challenges must be addressed in this context. One is reconciling integrated data types from autonomous laboratory systems to inform larger-scale models. The second is identifying and prioritizing which datasets must be generated by an autonomous laboratory system. As outlined in Ch. 4: Data Curation, Integration, and Products (p. 25), several data gaps and needs have been identified, primarily in a field context. This presents sub-challenges for prioritizing (1) how to address data gaps and needs in an autonomous laboratory system in a controlled environment (i.e., utilizing conventional laboratory assays) and (2) how to create field-deployable autonomous laboratory systems that could utilize networked field sensors tied to a portable field laboratory to modulate temporal sampling and assaying as needed.

Future AI Advances

Many ML approaches, such as ANNs, are already used routinely by the research community to resolve issues of missing data caused by instrument errors. Further incorporation of AI approaches into improved observational and experimental platforms (i.e., UAVs, edge-computing driven sensor platforms, and autonomous laboratories) can help resolve long-standing data gaps that have slowed progress toward a predictive understanding of the methane cycle and provide more robust benchmark datasets to evaluate and compare predictive model performance. These platforms require co-development with the computational infrastructure needed to support them (see Ch. 8: Computing Infrastructure, p. 49). AI methods can be used to improve data assimilation and sensitivity analysis (see Ch. 5: Multiscale Modeling, p. 33), and the output from those analyses can be used to design focused sampling campaigns, leading to more efficient data collection and scientific discovery.

A researcher uses a floating chamber and a gas analyzer to measure methane at Old Woman Creek. [Courtesy The Ohio State University]

4 | Data Curation, Integration, and Products

Diverse datasets are needed to close the methane budget, determine changes to methane cycling and fluxes due to climate change and extreme events, and identify methane mitigation measures. These objectives require data spanning multiple spatial (molecular to global) and temporal (hourly to multidecadal) scales (see Fig. 4.1, p. 26). They also require integration of interdisciplinary measurements such as genomic and phenotypic information to quantify biotic controls on methane production, consumption, and release; biogeochemical fluxes; climate and hydrological drivers of methane cycling; and anthropogenic emissions. The increasing availability of curated and integrated datasets has proven essential for advancing artificial intelligence (AI) and machine learning (ML) applications in many scientific domains.

The use of data-driven models requires easily accessible and well-documented datasets as inputs, as well as standardized benchmark data products for comparing model performance. However, challenges exist in acquiring, curating, and synthesizing relevant data from different sources for scientific use and AI/ML applications. First, datasets and databases that are relevant, available, and useful to address specific scientific questions need to be identified. Next, available data must be converted into model-ready data products. Conversion may require transforming existing data (e.g., by subsetting or scaling) to appropriate spatial domains and resolutions; harmonizing inconsistent variable names, units, and data formats; performing quality assurance and quality control (QA/QC); and gap-filling missing data.

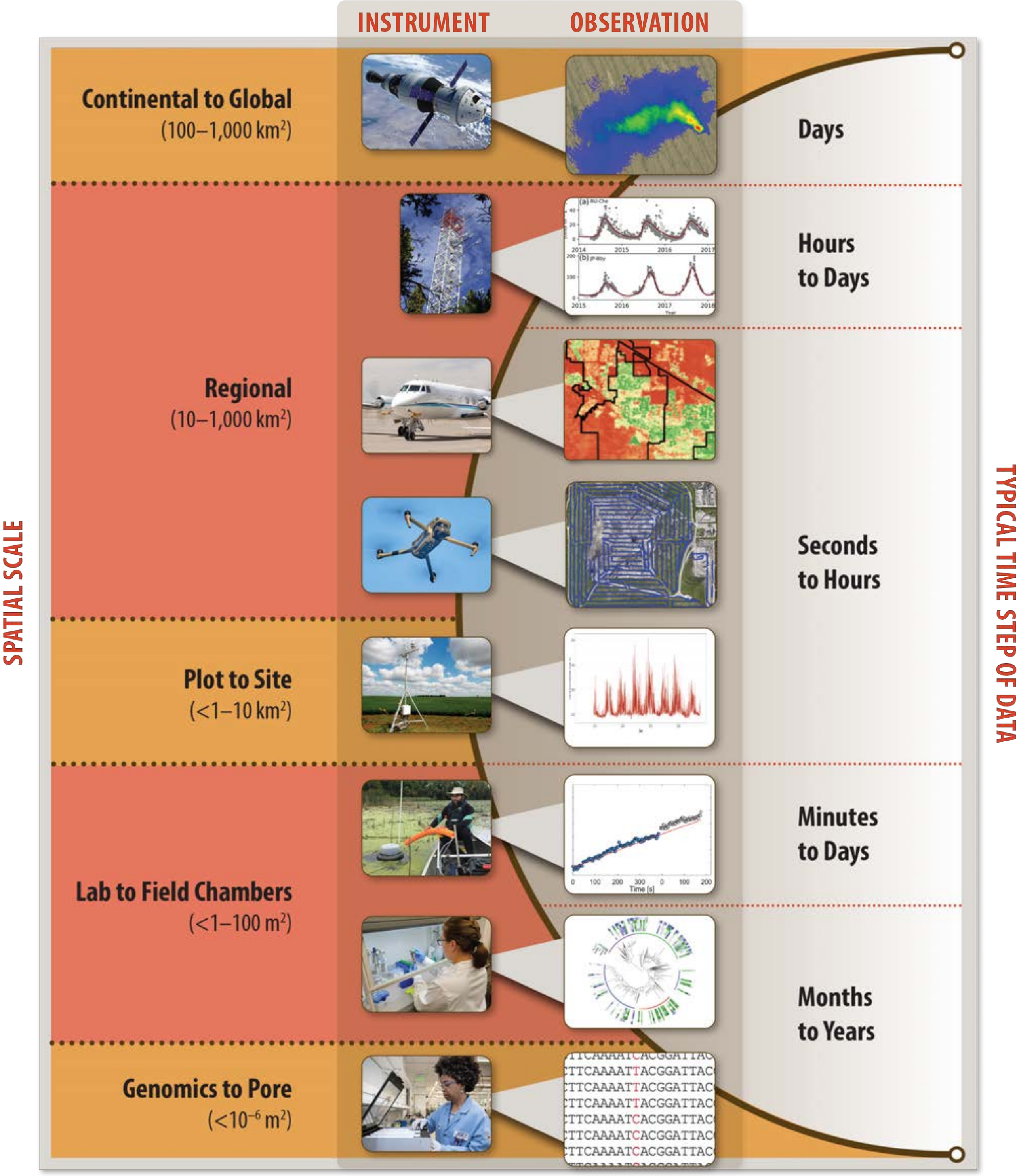


Fig. 4.1. Integrating Data Across Spatial Scales and Temporal Time Steps. Examples of instrumentation for collecting methane observations, along with associated output datasets, illustrate the inherent challenges of integrating data from different sources in space and time. The spatial scales reflect the typical footprint of each observational approach. The temporal timestep scale shows observation frequency in typical data products and represents approximate order of magnitude estimates with potential aggregation of high-frequency data. Instrumentation may collect data at higher spatial or temporal resolution, causing spatial scales and temporal timesteps to vary widely according to research goals and deployment strategy. [See Appendix E, p. 117, for image credits.]

The process of identifying and converting vast amounts of publicly available data into AI-ready products can be time-consuming and labor-intensive. Roughly 80% of processing time is spent preparing data and only 20% is spent analyzing data or modeling, prompting the term "the 80/20 rule of data science." However, the advent of new AI/ML technologies, especially large language models, offers opportunities to accelerate data discovery and integration and improve the productivity of scientific workflows.

This chapter describes how the general challenges outlined above lead to using AI/ML to advance predictive understanding of the methane cycle. The Data Gaps and Needs section, this page, describes data requirements for different scientific challenges identified in the workshop. The AI-Ready Data Synthesis and Benchmark Products section (see p. 28) then identifies opportunities to address these data needs by combining existing data into value-added products that can easily be applied to data-driven modeling. The Challenges in Data Availability and Curation section (see p. 30) outlines bottlenecks to making existing data usable for hypothesis testing, modeling, and analysis (e.g., data quality checks, uncertainty quantification, harmonization, and benchmark product creation). Finally, the Enabling Data Discovery and Integration with Artificial Intelligence and Machine Learning section (see p. 31) highlights challenges and opportunities associated with using AI/ML to make data scientifically usable, such as by increasing data discoverability across sources, improving reusability of existing data, and enabling data integration and scaling to appropriate resolutions.

Data Gaps and Needs

Vast amounts of Earth and environmental science data are now available for use in data-driven methane cycling models, including observations from remote sensing applications, sensor networks, metagenomic and transcriptomic analyses, and model simulation outputs. These data can be combined in myriad ways, so the first step toward using AI/ML to advance understanding of the methane cycle requires identifying data gaps and needs for specific scientific use cases.

A subsequent step is to determine whether existing data are available with sufficient spatiotemporal coverage, quality, resolution, and metadata for use in data-driven models.

AI4CH$_4$ workshop participants identified several gaps in the body of multiscale, multidisciplinary datasets available to accurately inventory top-down and bottom-up methane emissions, advance modeling and analysis, and improve predictive understanding of the methane cycle. Some data gaps can be filled by collecting new observations, synthesizing existing datasets, or running simulations to produce synthetic data. Scientific challenges and associated data needs relating to the methane cycle include:

- **Improving bottom-up estimates from wetlands**, which are the largest natural sources of methane. Quantification of wetland transport and emissions requires more reliable estimates of the global extent of wetlands and other waterbodies (see Stachelek et al. white paper, p. 72); synthesis of existing data, such as eddy covariance and chamber measurements (see Feng et al. white paper, p. 106; Yuan et al. white paper, p. 64); and additional data or observations to improve models that extrapolate bottom-up measurements to larger spatial scales.
- **Parameterizing and reducing uncertainties in process models used to estimate bottom-up emissions**, with direct biological and anthropogenic attribution of methane sources. Specifically needed at large spatial scales are high-resolution data for estimating plant-trait model parameters (e.g., vegetation type, leaf area index, and root traits including gas transport capacity) and, more broadly, ecosystem productivity (see Malhotra et al. white paper, p. 91). Data on plant and microbial trait distribution (e.g., genomic potential, abundance of methanogenic and methanotrophic pathways) and models of genotype-phenotype relationships could be used to improve such parameterization (see Ch. 5: Multiscale Modeling, p. 33).
- **Improving constraints on processes such as methanogenesis and methane oxidation** (e.g., Zhu et al. white paper, p. 62). Such advances

would require (1) co-located microbial community dynamics and flux data (e.g., Bueno de Mesquita and Tringe white paper, p. 70); (2) soil moisture and redox conditions and associated concentrations of oxygen or other electron acceptors; (3) dissolved and atmospheric methane isotope measurements; (4) insight into fine-scale methane cycling changes using imaging technologies, such as computed tomography scanning and neutron tomography, combined with geochemical and flux data (see Mayes et al. white paper, p. 97); and (5) the ability to integrate molecular- to plot-level microbial data into field- to global-scale ecosystem models (e.g., Mayes et al. white paper, p. 97; Xu and Rodrigues white paper, p. 80). Achieving these objectives is particularly challenging given the small scales and snapshot nature of these data. Also needed is partitioning of wetland and other biogenic fluxes into methane production and oxidation using existing data products such as FLUXNET-2015 (see Zhu et al. white paper, p. 62).

- **Quantifying the spatiotemporal heterogeneity of methane measurements**, particularly at monitoring sites with continuous, spatially extensive data. This can be achieved by co-locating instrumentation, synthesizing data from continuous chamber measurements at eddy covariance sites (see Yuan et al. white paper, p. 64), and decomposing flux data (see Chu et al. white paper, p. 77). Methane ebullition is a dominant methane release pathway from aquatic systems but is difficult to quantify due to high variability in space and time. Data are needed to capture the fluxes, spatial extent, and corresponding drivers (e.g., atmospheric pressure and water table depth) of bubbling and other episodic methane sources at large scales, particularly during triggering disturbance events such as storms, decreasing water levels in waterbodies, and wildfires (Varadharajan and Hemond 2012; Quebbeman et al. 2022; Zhu et al. 2022). New approaches to obtaining highly resolved spatiotemporal fluxes can help constrain the hot spots and hot moments of methane release. Such approaches include combining spatially extensive aerial or aquatic imaging (e.g., Berg et al. 2022; Chen et al. 2022; Keremedjiev et al. 2022) with temporally high-resolution measurements using, for example, AI-assisted autonomous instrumentation (see Ch. 3: Observations, Experiments, and Discovery, p. 19).
- **Quantifying anthropogenic sources of methane**, such as dairy farming, fossil fuel extraction and transportation, landfills, agriculture, and industrial emissions (see Morris et al. white paper, p. 89). This would require easier access to proprietary data spread across different sources; such data are inherently challenging to obtain. In addition, data on management practices (e.g., inputs, chemicals, and irrigation) for economically important crops and grazing lands in different agro-ecological regions would enable identification of methane mitigation measures.
- **Quantifying top-down estimates and aggregation across scales with higher spatiotemporal resolution** using remote sensing at different scales (e.g., satellite imagery and drones). Remote sensing can provide fine-resolution temporal and spatial surface characteristics, such as vegetation indices, surface temperature, and soil moisture, which are rarely available at monitoring sites (see Chu et al. white paper, p. 77).

In addition to multiscale measurements, data are needed across a variety of natural and managed ecosystems, including wetlands, peatlands, forests, croplands, lakes and reservoirs, streams, estuarine systems, oceans, tundra, and human-dominated systems (e.g., oil and gas fields and urban environments). A top priority is the need for co-located measurements in space and time at monitoring sites for use in models. This may require new measurements (see Ch. 3: Observations, Experiments, and Discovery, p. 19) as well as new data products after curation and integration of existing data (see AI-Ready Data Synthesis and Benchmark Products, this page).

AI-Ready Data Synthesis and Benchmark Products

To address the identified data needs, workshop participants highlighted opportunities for data synthesis efforts and products that provide training or validation data for AI and ML applications. AI-ready datasets are products that have been collected into a single harmonized format, curated (i.e., QA/QC, gap-filling, or scaling to appropriate resolutions), and linked to useful metadata. The research community has recently produced several integrated data products relevant to the methane cycle (see "Example Methane Datasets" sidebar, this page).

Opportunities exist to synthesize resources and datasets, particularly those supported through DOE funding and other U.S. federal agencies, into benchmark products for AI/ML applications. Potential targets include:

- DOE's AmeriFlux, global FLUXNET data, and other datasets generated as part of the AmeriFlux Year of Methane (ameriflux.lbl.gov/year-of-methane/year-of-methane).
- DOE's Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE): natural gradient measurements and manipulative experimental data from projects such as (1) Coastal Observations, Mechanisms, and Predictions Across Systems and Scales (COMPASS), (2) Spruce and Peatland Responses Under Changing Environments (SPRUCE), and (3) Next-Generation Ecosystem Experiments (NGEE) in the Arctic and tropics (ess-dive.lbl.gov).
- DOE's National Microbiome Data Collaborative (NMDC) multiomics microbiome data (microbiomedata.org).
- DOE's International Land Model Benchmarking (ILAMB) project: model-data intercomparison and integration datasets (ilamb.org/datasets.html).

Example Methane Datasets

BAWLD-CH4 is a synthesis dataset of small-scale, surface-methane flux data in boreal and Arctic regions from 540 wetland and non-wetland terrestrial sites and 1,247 aquatic sites (lakes and ponds) compiled from 189 studies (Kuhn et al. 2021).

COSORE is a community database of continuous soil respiration and other soil-atmosphere greenhouse gas flux observations (Bond-Lamberty et al. 2020).

FLUXNET-2015 is an AmeriFlux data product providing carbon dioxide, water, energy fluxes, and other meteorological and biological measurements from 212 sites (over 1,500 site-years, up to and including 2014) at eddy covariance towers around the world (Pastorello et al. 2020).

FLUXNET-CH4 is an open-source data product from eddy covariance towers. It consists of half-hourly and daily gap-filled and not gap-filled aggregated methane fluxes and meteorological data from 79 sites globally that span 42 freshwater wetlands, 6 brackish and saline wetlands, 7 formerly drained ecosystems, 7 rice paddy sites, 2 lakes, and 15 uplands (Knox et al. 2019; Delwiche et al. 2021). This dataset was used in a long short-term memory network (LSTM) and a modified causal LSTM to identify the primary drivers of wetland methane emissions (Yuan et al. 2022). It was also used to create an upscaled product (McNicol et al. 2023).

FRED is a global fine-root trait database (roots <2mm in diameter) consisting of more than 150,000 observations of more than 330 root traits, with data collected from more than 1,400 data sources (roots.ornl.gov).

Methane Working Group, of the Coastal Carbon Network, aims to compile all methane flux data from continental U.S. coastal habitats (not mangroves) to parameterize and validate process-based methane models (serc.si.edu/coastalcarbon/data).

TRY Database is an open global dataset of curated plant functional traits (try-db.org), which are standardized and quality checked. This dataset integrates more than 700 published and unpublished datasets.

- DOE's Molecular Observation Network (MONet): database of molecular-level and microstructural information on the composition and structure of soil, water, resident microbial communities, and biogenic emissions (see Karra et al. white paper, p. 108; www.emsl.pnnl.gov/monet).
- National Science Foundation's National Ecological Observatory Network (NEON): terrestrial, aquatic, atmospheric, and remote sensing data and samples (nsf.gov/news/special_reports/neon).
- International Soil Carbon Network's community platform for communicating, modeling, and sharing data to advance science questions related to soil organic matter and soil organic carbon (iscn.fluxdata.org).

A comprehensive effort to synthesize data would require collaborations with multiple agencies and international entities (e.g., NASA, DOE Office of Fossil Energy and Carbon Management, and United Nations International Methane Emissions Observatory). Examples of data product and synthesis needs that address the science challenges described in Data Gaps and Needs (see p. 27) include:

- High-resolution historical and predictive maps of methane fluxes and uncertainties for different ecosystems, including global wetland methane emissions with eddy covariance and chamber measurement data.
- Soil flux database for methane.
- Fine-scale maps of synthesized landscape properties, such as topography (e.g., for identifying peatland distribution and complex hummocks and hollows), water table depth, vegetation, land use, and land cover (including wet body extents) from new satellite constellations (e.g., PlanetScope and Hydrosat) to obtain fine spatiotemporal surface characteristics.
- Synthetic database of methane-related biogeochemical variables generated by mechanistic ecosystem and microbial models.
- Reconciled observations at different scales from multiple monitoring sites and synthesis of methane fluxes from eddy covariance towers, drones, and satellites with *in situ* geochemical, omics, and other ancillary data.
- Synthesized dataset of anthropogenic emissions.

Challenges in Data Availability and Curation

Enabling reuse of existing data and creation of AI-ready datasets to answer new science questions requires (1) providing sufficient metadata outlining the purpose and methods of data collection, descriptions of variables and data processing, and data authorship and use guidelines; (2) adopting standardized formats for reporting metadata and structuring data files; (3) publishing data in open-access repositories that support free and fair use data licenses; and (4) providing connected and searchable infrastructure to enhance data findability (see Ch. 8: Computing Infrastructure, p. 49). The scientific community has recently moved toward publishing open data in repositories and has adopted FAIR principles, which improve data findability, accessibility, interoperability, and reusability (Wilkinson et al. 2016). This shift has enabled significant amounts of scientific data to become available for data-driven modeling and for comparisons between process-based and ML model outputs.

However, challenges remain in adopting FAIR principles and creating data products that drive AI/ML models for methane cycling. First, multiscale, multidisciplinary datasets are spread across data repositories and sources (see Data Gaps and Needs, p. 27). Some of these are open access, such as AmeriFlux, ESS-DIVE, DOE Systems Biology Knowledgebase (KBase), NEON, NASA's Distributed Active Archive Centers, NASA's Carbon Mapper, NMDC, and GenBank, but others are proprietary with restricted access or differing usage policies. Furthermore, other data, such as fine-scale imaging data, may lack a dedicated repository and thus consistent metadata or file formats, resulting in fragmented datasets that are difficult to combine. Finally, large datasets (e.g., numerical model outputs) are not widely archived due to size limitations of data repositories, which creates a bottleneck to using these data to train emulator models or compare model performance.

Available reporting formats that enable standardized data and metadata reporting include:

- AmeriFlux file format and metadata for eddy covariance measurements (Pastorello et al. 2020).
- ESS-DIVE soil respiration, amplicon, and other environmental data and metadata reporting formats (Bond-Lamberty et al. 2021; Crystal-Ornelas et al. 2022a).
- Genomic Standards Consortium's standards for metagenomic measurements, including MIxS, MIGS, MIMS, and MIMARKS (Yilmaz et al. 2011).
- ESS-DIVE archiving guidelines for terrestrial modeling data (Simmonds et al. 2022).
- Persistent identifiers (e.g., International Generic Sample Number) with relevant metadata on locations (Crystal-Ornelas et al. 2022b) and samples (Damerow et al. 2021) to link related datasets and track co-located measurements from samples split for different laboratory analyses.

Ontologies or standardized vocabularies for key variables could facilitate data synthesis and comparison of many relevant datasets that still lack domain-specific reporting formats, including chamber and bubble trap measurements; methane-specific omics metadata; fine-scale imaging and synchrotron measurements; and monitoring, reporting, and verification (MRV) for tracking anthropogenic emissions. In addition, instituting conventions for reporting data and metadata for different measurements could enable interoperability across systems. Centralized QA/QC procedures that leverage ML methods (see AI-Ready Data Synthesis and Benchmark Products, p. 28) can enable consistent, scalable data pre-processing prior to use. Finally, data for mechanistic and ML models may require pre-processing, gap-filling, and uncertainty quantification (see Ch. 6: Data–Model Integration and Benchmarking, p. 39).

To accelerate the availability of reusable data, a broader effort among methane researchers is needed to encourage open data release of observational and modeling datasets and to adopt existing standards, ontologies, or develop new ones. Integrating standardized workflows and enabling infrastructure (see Ch. 7: Enabling Data and Model Exchange, p. 45; Ch. 8: Computing Infrastructure, p. 49) into the research and data collection lifecycle would also lower the barrier to data curation. Developing tools that make it easier for contributors to provide well-curated and standardized data are needed. Scientific data contributors and managers need sufficient resources and incentives to support data management efforts for curation, standardization, QA/QC, pre- or post-processing, and publication. Finally, solutions and guidelines to archive model outputs, from both process and ML models, are needed.

Numerical model output datasets can be large and grow over time and, in many cases, it will not be possible to store entire model outputs. Yet using surrogate models to emulate computationally expensive process models is growing, so sufficient amounts of simulation data with different parameterizations must be made available to train such models. Potential research avenues include using data compression techniques for model output and techniques to identify representative model runs for archiving from a large number of ensemble outputs.

Enabling Data Discovery and Integration with Artificial Intelligence and Machine Learning

Current data discovery and integration workflows are typically bespoke and tailored for specific applications. Opportunities exist for AI and ML to enhance data searches, curation, and integration for Earth sciences. A comprehensive overview of this topic is outlined in the "Data Acquisition to Distribution" chapter and several domain-specific chapters of the AI4ESP report (U.S. DOE 2022). Examples of successful applications include anomaly detection or denoising methods for QA/QC (Blázquez-García et al. 2020), various methods for imputation (e.g., Mital et al. 2020; Ryu et al. 2020; Park et al. 2023), and downscaling remote

Opportunities exist for AI and ML to enhance data searches, curation, and integration for Earth sciences.

sensing and other spatial data to desired resolutions (see Ch. 6: Data–Model Integration and Benchmarking, p. 39). Data integration tools like BASIN-3D enable more automated, on-demand synthesis of time series data without continual updates to data products. Such tools have been used to integrate data for ML models (Varadharajan et al. 2021).

More recent opportunities for potential exploration by the data management research community involve large language models (LLMs) such as GPT4 (openai.com/gpt-4) and Mistral (nlp.stanford.edu/mistral) to augment existing metadata, enable data discovery and synthesis (Fernandez et al. 2023), and enable interrogation of data and metadata via natural language interfaces. These disruptive technologies have great potential to extract information from unstructured text (e.g., abstracts and publications) as well as files that may be structured in different ways across datasets.

Currently, data synthesis and search across databases requires domain expertise to manually map data with different variable names, units, and collection methods into standardized terms. However, recent AI approaches have demonstrated the ability to link ontologies and resolve semantic differences (e.g., Toro et al. 2023). Such approaches can enhance data discovery, particularly when paired with tools like LinkML (Moxon et al. 2021; linkml.io), which can describe and link related datasets.

While the use of LLMs in genomics is growing (Tang 2023 and references therein), their use in Earth sciences is nascent and has primarily been demonstrated for keyword classification to improve searches (Ramachandran et al. 2022). Additional applications include improved semantic searches, automated taxonomy classification, and text summarization. The ultimate potential for LLMs is to enable knowledge discovery.

Remaining questions to be resolved regarding LLM application include finding the best approaches for model building, fine-tuning, or training and identifying architectures for augmenting LLMs with up-to-date knowledge (e.g., Retrieval Augmented Generation; Lewis et al. 2021). LLM reliability and trustworthiness presents an issue, with current technologies suffering from so-called "hallucinations." In the BER research space, pairing LLMs with the model-experiment (ModEx) approach of iterative experimental testing is crucial to building scientific understanding. Improved approaches for determining LLM accuracy and mitigating potential bias are needed (Dentella et al. 2023). LLM use is also limited by the challenges identified earlier (see Challenges in Data Availability and Curation, p. 30), including a lack of benchmark datasets, inadequate standards, and insufficient adoption of existing standards in the Earth sciences. Therefore, a focus on developing tools and other solutions to support and incentivize data generators to adopt community formats and ontologies, as well as curate the data they make available (see Ch. 7: Enabling Data and Model Exchange, p. 45), is essential to advancing AI services that further accelerate data discovery and knowledge generation.

Researchers discuss simulations of watershed biogeochemistry. [Courtesy Lawrence Berkeley National Laboratory]

5 | Multiscale Modeling

Estimating methane feedbacks to the Earth system requires a predictive modeling framework for methane fluxes at multiple scales ranging from interacting microbial populations to continental expanses. However, the complexity and variability of terrestrial methane sources and sinks have long posed significant challenges to predictive modeling and accurate forecasting. Traditional modeling methods often struggle to capture heterogeneous and dynamic methane processes in terrestrial ecosystems arising from non-linear and scale-emergent processes, such as identifying hot spots and hot moments for methane release (Sturtevant et al. 2016). Accurately simulating methane fluxes also requires predicting numerous above- and below-ground processes and their complex interactions. Recent artificial intelligence (AI) advances may greatly improve the capacity to model methane dynamics and improve quantification of contributing processes. AI can be leveraged to capture higher-order relationships among variables, which improves prediction accuracy across scales.

Advancing multiscale predictive modeling frameworks requires integrating top-down and bottom-up modeling and observation methods, which are markedly different for methane (see Fig. 2.1, p. 10). Top-down methods typically rely on atmospheric methane concentration measurements from towers, satellites, or airplanes. Inverse atmospheric transport modeling is then used to estimate methane fluxes, such as from wildfires, waterbodies, or wetlands. Given the often sparse spatial and temporal distribution of these concentration measurements, top-down methods constrain modeling to coarse spatial scales. Additionally, they provide only budget-level emissions estimates, from which it is difficult to gain mechanistic insights.

In contrast, bottom-up predictions are generally driven by mechanistic models that may provide high spatial and temporal resolution for both methane emissions and associated processes. However, bottom-up model predictions often vary significantly from each other, likely due to large parametric and structural uncertainties. Structural uncertainties stem from different mechanistic hypotheses, resulting in multiple possible model algorithms that may be used to represent the same process. Mechanistic model structures and parameters may be constrained by field observations (e.g., flux chamber or eddy covariance), but in practice this type of model-data synthesis has proven difficult due to model complexity and computational expense.

> *AI methods will likely provide advantages over traditional data assimilation techniques by reducing the required model ensemble size and enabling much faster computation.*

The current lack of efficient model-data synthesis methods presents a major barrier to advancing methane flux prediction from site to global spatial scales and from hourly to decadal timescales. Ideally, multiple observation types and scales could be assimilated into land-surface models simultaneously, including both top-down estimates and field observations. AI methods will likely provide advantages over traditional data assimilation techniques by reducing the required model ensemble size and enabling much faster computation.

Data Assimilation and Sensitivity Analysis

AI methods may greatly accelerate data assimilation and overall model-experiment (ModEx) workflows. For example, a common method for calibrating physical model parameters is a Bayesian technique known as Markov Chain Monte Carlo (MCMC). In addition to parameter optimization, MCMC also quantifies parameter and prediction uncertainty given a set of observational constraints. MCMC typically requires tens of thousands or more serial model evaluations depending on the number of uncertain parameters and the complexity of model behavior. Therefore, it is generally computationally infeasible to apply MCMC directly to complex land-surface models that simulate methane fluxes along with many other interacting terrestrial processes. Instead, model surrogates may be developed by performing ensemble simulations in parallel and training AI models on the output.

Surrogate models, sometimes referred to as emulators, predict selected outputs of interest from the original model as a function of parameters or other inputs. They are developed by fitting functions to model ensembles. Traditionally, polynomial functions, radial basis functions, or Gaussian process models have been used to develop surrogates (e.g., Sargsyan et al. 2014; Müller et al. 2015) but AI methods, like deep neural networks (DNNs) and long short-term memory (LSTM) networks, are also promising (see Feng et al. white paper, p. 106; Ricciuto et al. white paper, p. 95). Surrogate modeling methods with increased accuracy improve predictions and efficiency because they require a smaller ensemble of the original model. While effective at the point scale, traditional methods have struggled with developing spatially and temporally resolved surrogate models, especially at high resolution. Dimension reduction approaches use high spatial and temporal autocorrelation in model outputs and have been successfully applied in combination with DNNs to generate surrogate models based on land-surface model ensembles (Lu and Ricciuto 2019; Dagon et al. 2020). Convolutional neural networks (CNNs) and autoencoders have the potential to further advance these capabilities.

Sensitivity analysis is another critically useful tool to quantify uncertainty, better understand model behavior, and guide observational campaigns (e.g., timing and placement of methane sensors in specific ecosystems). Land-surface models often contain dozens, if not hundreds, of uncertain parameters and constraining them simultaneously is unfeasible. Instead, global sensitivity analysis (GSA) can identify a subset of uncertain model parameters that are important for addressing a particular scientific question or model output of interest. For example, certain model parameters may be more sensitive during simulated periods of high ebullition, suggesting that measuring these ecosystem properties may improve prediction of such events. Riley et al. (2011) developed the CLM4Me model and computed the sensitivities of methane parameters, finding that they were different among different ecosystems.

To better understand the impact of methane parameter uncertainty on the Earth system, parameters must be analyzed in combination, including from non-methane processes. However, such high-dimensional GSA is computationally expensive and generally requires numerous model evaluations to determine key parameter sensitivities and interactions. As with model calibration, surrogate modeling is highly useful in GSA.

Advancing Predictive Capabilities

Traditional modeling approaches have a limited ability to incorporate microbial mechanisms controlling methane production and consumption into large-scale models, including Earth system models. In addition to the data challenges associated with large differences in measurement scales (see Ch. 4: Data Curation, Integration, and Products, p. 25), high computational costs pose challenges to capturing empirical observations, constraining model parameterization, and quantifying modeling uncertainty. Determining the level of complexity needed in microbially explicit models is difficult and a continuing research focus (see Song white paper, p. 99; Xu and Rodrigues white paper, p. 80).

AI approaches offer several potential solutions:

- **Microbial Modeling.** Classifier models can enable phenotypic predictions based on genomes, proteomes, transcriptomes, metabolomes, or other cellular activity proxies, which can be validated in culturable organisms (see Chapter 3: Observations, Experiments, and Discovery, p. 19). They can also overcome certain scaling issues (e.g., limited knowledge of gene annotations and metabolic pathways) and produce higher-quality mechanistic models (Kavvas 2020; Bi et al. 2023; Liu et al. 2023). These models can then improve parameterization of multiscale methane models for simulations within an Earth system modeling framework. AI approaches like artificial neural networks (ANNs) have been used to predict microbial community dynamics and activity (e.g., Larsen et al. 2011). Meanwhile, researchers are exploring DNN methods, including LSTM, to generate surrogate models for integration into coarser-scale models (see Oh et al. white paper, p. 74).
- **Surrogate Models.** In addition to their use in model sensitivity analysis and calibration, surrogate models can replace expensive computations in specific parts of the model codebase. Data-driven AI models may also replace computationally complex or poorly represented processes. Such hybrid approaches are increasingly used in Earth system models, which contain subroutines describing different land-surface processes. The land model component of DOE's Energy Exascale Earth System Model (E3SM), for example, contains more than 200,000 lines of code in over 100 subroutines to describe biogeochemical and biogeophysical processes. A DNN-based surrogate model for wildfire, developed by Zhu et al. (2022), simulates a burned area with 90% accuracy compared to observations and significantly reduces the number of model parameters. This model was embedded within E3SM's Land Model (ELM), creating a hybrid version that includes both machine learning and mechanistic submodels. This approach may be especially valuable for investigating wildfire impacts on methane emissions by reducing uncertainties within the wildfire model itself, and by enabling more detailed investigation of uncertainties in the mechanistic methane submodel. Similar approaches may be integrated into other multiscale modeling frameworks, enabling the representation of fine-scale processes within coarse-scale land-surface models.
- **Remote Sensing and Ground-Based Observations.** These measurements should be combined to calibrate multiscale models and to develop hybrid modeling approaches. For example, the FLUXNET database, which provides increasingly

rich energy, carbon dioxide, and methane flux data from diverse ecosystems, is already used in various AI approaches, such as developing reliable evapotranspiration predictions (ElGhawi et al. 2023). Knowledge-guided machine learning, a hybrid approach in which physical constraints and mechanistic response functions are incorporated into cost function or training, is a promising method for improving the predictability of methane cycling based on FLUXNET or other observation networks. While not all eddy covariance towers measure methane, FLUXNET-CH4 includes over 80 sites globally. Although global coverage is sparse in tropical regions, it covers boreal, temperate, and Arctic regions reasonably well (Delwiche et al. 2021). Eddy covariance networks may be augmented by chamber-based measurements that have smaller footprints but may be more numerous and better suited to capturing localized hot spots and hot moments. Chamber measurements of methane have been synthesized in several recent studies (e.g., Guo et al. 2023), but more are needed. The Continuous Soil Respiration (COSORE) database (Bond-Lamberty et al. 2020), while initially developed for heterotrophic respiration measurements, also accommodates methane measurements and provides data in standardized formats that could be readily used by machine learning approaches. Satellite measurements will be critical for regional scaling, including new launches such as MethaneSAT that will have the capability to estimate fluxes at high spatial resolution. AI can also greatly improve land use classification for remote sensing (Bourgeau-Chavez et al. 2021; Rodriguez et al. 2023), such as providing more accurate areal extent of wetlands, peatlands, forests, and lakes.

- **Real-Time or Near-Real-Time Model Data Assimilation.** This capability—enabled by edge computing, intelligent sensors, and advanced model-data integration techniques—can greatly accelerate the ModEx cycle. AI-driven models can guide measurements and automate experiments. In particular, these approaches could greatly increase the volume of data collected during sudden disturbances, extreme events, or hot spots and hot moments that strongly impact methane fluxes and occur with little lead time. Integrating these data into models will then improve the predictability of future events. AI approaches may also be applied to check for errors in data (e.g., sensor bias or drift) that would negatively impact model predictions. These approaches may be especially useful for capturing methane emissions from water bodies, which can experience dynamic fluctuations in size over short timescales that can contribute disproportionately to methane emissions (Pi et al. 2022). A near-real-time forecasting capability was developed in the Spruce and Peatland Reponses Under Changing Environments (SPRUCE) whole-ecosystem warming and elevated carbon dioxide experiment in a northern Minnesota bog (Huang et al. 2019). Although information from this study did not directly inform sensors, next-generation ecosystem experiments could incorporate these capabilities into hardware and software designs.

Challenges and Opportunities

An often-mentioned limitation to implementing AI-based predictive frameworks is the “black box” nature of the underlying approaches, and while machine learning model interpretation of individual processes is challenging, significant progress is underway. For example, LSTM networks can reveal the importance of driving variables and their time dependencies (Lu et al. 2022). The dependencies identified in an LSTM are based on correlations rather than causality, so interpretability may be limited due to confounding variables. However, LSTMs can be constrained to better infer underlying causal relationships. Such an approach was used to better understand and predict methane fluxes over different wetland types using FLUXNET observations (Yuan et al. 2022).

Managed systems present further challenges to modeling methane emissions. Processes currently parameterized for natural systems perform poorly in agricultural areas and other managed systems. This challenge impedes global predictability of anthropogenic methane, of which a large proportion comes from rice cultivation and other human management practices. Land management practices, such as flooding and draining of fields for cultivation, are therefore crucial to include in models. Impacts of specific management practices may be assessed by eddy covariance methods (e.g., Runkle et al. 2019) and then incorporated into data assimilation or hybrid approaches. Wetland draining or restoration may also strongly impact methane emissions. Multisector dynamics models, such as the Global Change Assessment Model (GCAM), can predict management decisions in the energy–human–climate system as a function of climate and socioeconomic factors and are the best available tools for assessing different policy scenarios.

Future AI Advances

In addition to multisector dynamics models, foundation models offer a promising avenue to enhance methane emission predictions from both natural and managed systems. Foundation models often comprise billions or even trillions of parameters, which enable them to encapsulate a vast amount of information and handle large datasets (e.g., high-resolution model outputs). Once trained on a diverse range of multiscale data, these models can be fine-tuned to specific tasks with smaller datasets. They may then be repurposed across different domains with fewer data requirements. Some foundation models are designed to process multiple types of data simultaneously, which can be particularly useful when integrating various data sources for methane prediction across scales, from chamber measurements to satellite imagery.

Foundation models (e.g., ClimaX) have recently been applied to numerical weather prediction, improving medium-range synoptic-scale forecasting for many atmospheric variables (Nguyen et al. 2023). Such models can accurately downscale climate model projections, potentially enabling better prediction of methane flux heterogeneity. More broadly, foundation models have demonstrated an ability to generalize across tasks and domains, meaning they can apply knowledge learned in one context to a related context. These models convert input data into high-dimensional representations (i.e., embeddings) which capture intricate patterns and relationships.

Such models could be used to learn the relationships between climate patterns and either simulated or observed methane fluxes at high temporal and spatial resolutions. Foundation models have the capacity to handle large volumes of information, which would be needed as land-surface models rapidly advance to higher resolution. Some foundation models could also incorporate new information into their existing knowledge base with minimal retraining, which would be vital for adapting to the evolving nature of datasets in real-world scenarios. Researchers and domain experts could interact with these models, refine outputs, and iteratively improve predictions. The architecture of foundation models, especially transformers, is inherently flexible, allowing them to be adapted to a wide variety of tasks beyond their initial training purpose.

Artist's rendering of big data analysis using artificial intelligence. [Courtesy Adobe Stock]

6 | Data–Model Integration and Benchmarking

New scientific knowledge and insights come not only from observational data but also from data integration with models. Such data–model integration can be accomplished in many ways, including developing data for model input or data-driven model training, assimilating data into models as boundary conditions or parametric constraints, verifying model development through data comparisons, and assessing model or multimodel fidelity by comparing model output with observational and reanalysis datasets using comprehensive benchmarking and diverse statistical metrics. As described in other chapters, measurement and observational data are required for integrating and constraining models across multiple spatial scales, from microbial and plant genomes to *in situ* ecosystem measurements, and from observations across watersheds and continents to the global scale (see Fig. 4.1, p. 26). Such measurements are being collected at varying temporal scales, ranging from a single time point to near continuous automated sensor measurements; however, in most cases, only subsets of the collected data are suitable for model integration due to sampling biases, insufficient sampling frequency or density, or inadequate model representation of targeted processes. In addition, finding and accessing appropriate data and relevant models that address a given science question remain enormous challenges.

These model and data integration challenges were addressed in the AI4CH$_4$ workshop white papers (see Appendix C, p. 60) and in thematic discussion sessions during the workshop. Workshop participants exchanged ideas for advancing methane research through large-scale data analytics and data-driven modeling, complementing traditional approaches as well as developing and applying machine learning (ML) models that rapidly advance the research enterprise for DOE and the broader science community.

Data Quality Control, Gap-Filling, and Synthesis

Large uncertainties associated with measurement gaps pose critical challenges to resolving top-down and bottom-up methane emissions estimates. For example, ground-based *in situ* measurements, common in investigations of methane-related processes, are often limited to easily accessible sampling locations and frequently contain temporal gaps due to technical and logistical issues. In addition, methane emissions measurements from soils, wetlands, and water bodies are available from a limited number of eddy covariance towers at wetland sites (e.g., FLUXNET-CH4; Delwiche et al. 2021), chamber measurements, and *in situ* measurements, and they are difficult and expensive to acquire. Such sampling challenges often result in missing or sparse hydrological, microbial, and plant trait data.

ML approaches are commonly employed to address these challenges by determining the representativeness of measurements, designing optimal sampling networks (e.g., Hargrove et al. 2003; Schimel et al. 2007; Keller et al. 2008), filling spatial or temporal sampling gaps (e.g., Morin et al. 2017), and interpolating and extrapolating spatial and temporal data even in highly heterogeneous environments (e.g., Kumar et al. 2016; Jung et al. 2019; Irvin et al. 2021; McNicol et al. 2023; see Fig. 3.1, p. 20). For example, since resource and logistical constraints often limit the frequency and extent of environmental measurements of carbon (i.e., methane and carbon dioxide), water, and energy fluxes in the Arctic, Hoffman et al. (2013) used a quantitative statistical methodology to stratify sampling domains. This approach informed sampling site selection and determined the representativeness of sites and networks across Alaska. In turn, the results informed future sampling site selection and maximized sampling across critical environmental gradients.

Using a similar technique, Pallandt et al. (2022) quantified the representativeness of a network of pan-Arctic eddy covariance sites to optimize planning for future network enhancements. One unique component of their optimization analysis was that some of the sites measured carbon dioxide in addition to methane fluxes, and some sites collected data only during the summer season rather than the entire year. Such operational inconsistencies must be considered in any comprehensive assessment of measurement representativeness.

To upscale FLUXNET eddy covariance observations of carbon dioxide, water, and energy fluxes to the global scale, Jung et al. (2011) used a ML technique called model tree ensembles (MTE). They trained MTE to predict site-level gross primary productivity, terrestrial ecosystem respiration, net ecosystem exchange, latent energy, and sensible heat based on vegetation indices, climate and meteorological data, and land use information. A similar approach could be applied to extrapolate methane fluxes to the global scale.

To integrate soil organic carbon (SOC) measurements across 2,374 soil profiles in the northern circumpolar region and to produce a spatially extant map of SOC, Mishra et al. (2020) combined multiple ML methods (i.e., gradient boosting machine, multi-narrative adaptive regression spline, support vector machine, and random forests). The researchers found that an ensemble prediction approach (see “Types of AI” sidebar, p. 3) employing multiple ML methods usually offered better predictions of observed SOC spatial variation, which strongly correlated with methane emissions in wet regions, than any single method alone.

Data Assimilation and Hybrid Modeling

A lack of long time-series methane emissions data limits the ability to initialize and run data-driven models using data assimilation approaches. Disparate data at different scales are stored at different data centers in different ways, creating challenges for data integration and synthesis (see Ch. 4: Data Curation, Integration, and Products, p. 25). However, new measurements and compilations of existing data (e.g., FLUXNET-CH4) could drive development of data-driven models and improve understanding of key processes and controls in the global methane cycle.

To characterize the performance of global freshwater wetland methane emissions models, Zhang et al. (2023) applied a wavelet spectral analysis to identify the dominant time scales contributing to model uncertainty in the frequency domain. They developed a Monte Carlo approach to incorporate flux observation error to avoid misidentification of time scales that dominate model error. Among their results, they showed that most models could capture methane variability at monthly and seasonal time scales for boreal and Arctic tundra wetland sites, but the models contain significant variability bias at seasonal time scales for temperate and tropical/subtropical sites. The work suggests a need to accurately replicate freshwater wetland methane flux variability, especially at short time scales, in future wetland methane model developments.

Yuan et al. (2022) used causality-guided ML to identify the importance of soil temperature across different wetland types. They captured a lagged response to ecosystem respiration and gross primary productivity in a subset of these wetland types and identified the importance of these factors in driving methane emissions at the sub-seasonal scale.

McNicol et al (2023) developed a six-predictor, random-forest, upscaling ML model (UpCH4) trained on 119 site-years of eddy covariance methane flux data from 43 wetland sites in the FLUXNET-CH4 Community Product to demonstrate model skill in producing realistic extra-tropical wetland methane emissions estimates. Such estimates will improve with more flux data. Near-term, essential sources of additional data will come from new satellite constellations, such as PlanetScope (planet.com) and Hydrosat (hydrosat.com), which provide very fine spatiotemporal surface characteristics (e.g., Moon et al. 2022).

Another challenge lies in connecting microbial community information to model-relevant functional information (e.g., metabolic rates). Most sequence-based datasets (e.g., 16S ribosomal RNA sequences, metagenomes, metatranscriptomes) provide information only about relative abundances of genes and organisms, not absolute abundances, making it infeasible to tease out complex microbe–environment relationships. Metagenomic, and especially metatranscriptomic, data paired with methane measurements are rare. Many more 16S datasets are available, but accurately predicting function and activity from these data presents a sizable challenge (Øyås et al. 2024; see Ch. 2: State of the Science, p. 9).

Data platforms, such as the DOE Systems Biology Knowledgebase (KBase), that combine genomic databases with analysis tools to compare genomes and construct genome-enabled models (GEMs) of microbes and microbial communities can facilitate development of both testable hypotheses and ML models of genotype–phenotype–environment relationships. After validation of these models in laboratory and field settings, they may serve as surrogates of complex biology in larger-scale models (see Ch. 5: Multiscale Modeling, p. 33).

Methane cycle process models could be enhanced by including ML-based parameterizations to produce hybrid models that combine process-based algorithms with ML-based modules. Such hybrid models can improve accuracy over traditional process-based models, and they consistently yield better performance on modern supercomputers. A hierarchy of hybrid models that employ common application programming interfaces (APIs) could enable reuse and interoperability of simulation code.

Model Benchmarking and Analysis

Growing model complexity, particularly with hybrid models, necessitates new methods for characterizing model performance and suitability. A systematic approach is needed to identify strengths and weaknesses of both process-based and ML models and to continuously verify that new model algorithms produce desired results (Randerson et al. 2009). Multimodel comparisons with observational data products, often called model benchmarking, provide useful information about which combinations of process representations yield better performance with respect to observational reference data. Model benchmarking tools, such as the International Land Model Benchmarking package (see Fig. 6.1, p. 43), have proven useful for characterizing model functionality and reproducibility, relative model performance, and fidelity of new model versions (Collier et al. 2018).

Model benchmarking and traditional analysis techniques are leveraging technology advances to improve scientific productivity. Interactive analysis of model output and observational data are increasingly performed with Python or R languages in Jupyter Notebooks, often on JupyterHub nodes co-located with storage systems that host and manage increasing data volumes. Commercial cloud vendors have helped propel this approach to data science, and cloud-based data storage technologies and dynamic provisioning of computational capacity are materializing even in traditional high-performance computing centers, including those operated for DOE.

Advancing Understanding of the Methane Cycle Using AI/ML

Achieving new insights into integrated process understanding across BER's research enterprises requires advancing infrastructure that enables aggressive use of both large data and complex models. Relevant processes operate at scales from the genome to the global Earth system, spanning at least 12 orders of magnitude.

New artificial intelligence (AI) and ML technologies, including large language models and foundation models, combined with new computational and storage infrastructure, offer opportunities to significantly advance information harvesting from data, drive models directly with data, and conduct very-large-scale data analytics. Creating new synthesized benchmark datasets for ML model training is crucial to enabling further progress in methane-related research. Such datasets should include long time-series measurements of methane fluxes from tropical and subtropical wetlands, genomic characteristics of methanogens across global wetlands, and globally distributed eddy covariance measurements of methane and related environmental drivers (e.g., meteorology and site characterization data). Associating microbial characteristics with functional responses will enable creation of new microbially explicit soil–carbon dynamics models that can underlie simulation experiments for the entire globe, and characterize microbial responses to climate change that will play a large role in future atmospheric methane levels and climate feedbacks.

Realizing these improvements to data and model accessibility will require a foundation of integrated research infrastructure (i.e., both hardware and software) for data management, AI/ML model building, simulation, and analysis at scale (see Ch. 8: Computing Infrastructure, p. 49). Similar investments in ecosystem- and smaller-scale experiments will be needed to fill data gaps.

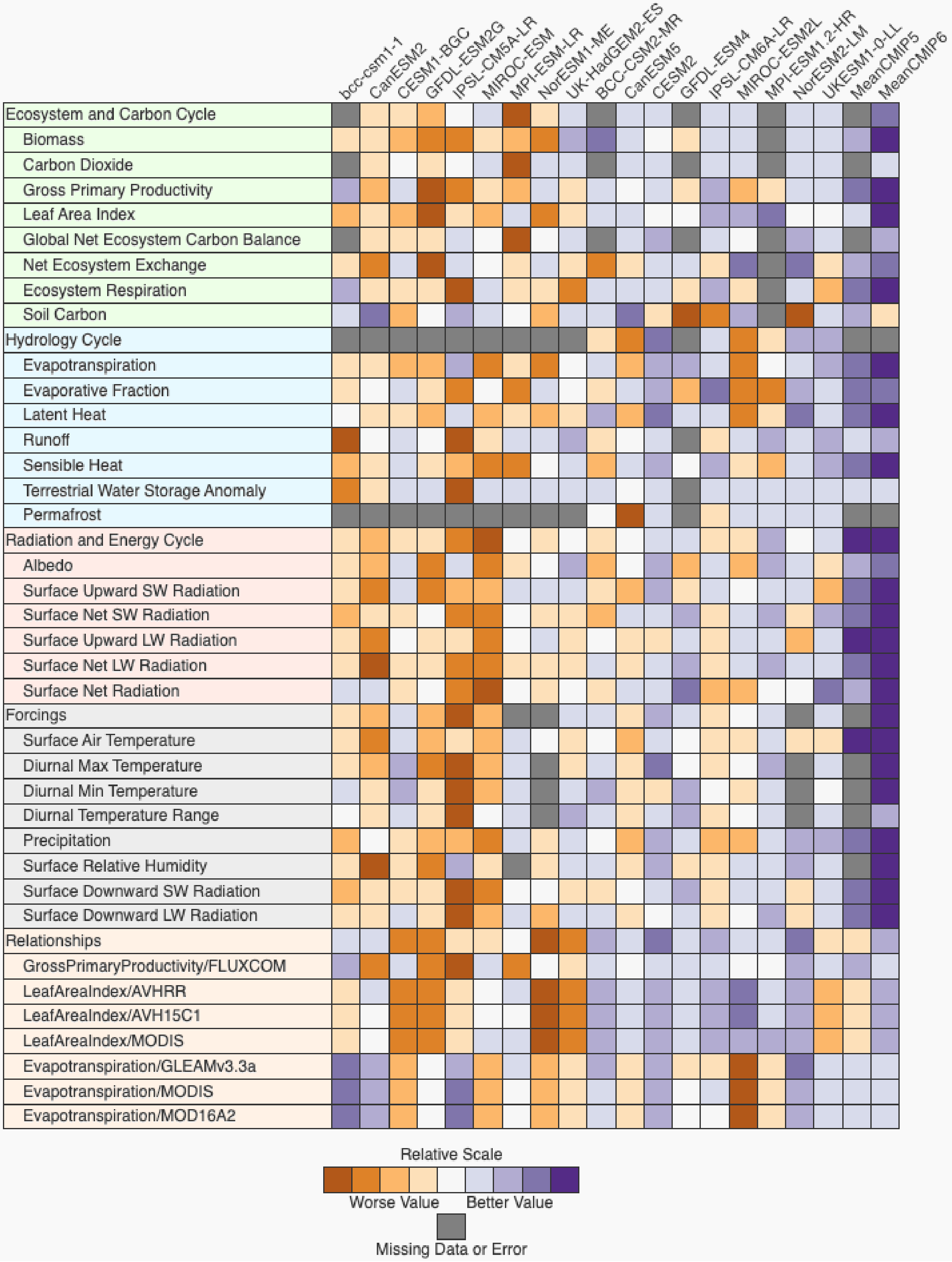


Fig. 6.1. A Tool for Intercomparing Complex Models with Data. This example portrait plot from the International Land Model Benchmarking (ILAMB) model-data intercomparison and integration package compares the performance of two generations of nine land surface models (labeled at top of each column) used in the fifth (left) and sixth (right) phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6), as well as the mean of the CMIP5 and CMIP6 models, for multiple variables and functional relationships (rows). Such comparisons characterize model functionality and reproducibility, relative performance, and fidelity of new model versions. [Courtesy of the RUBISCO Science Focus Area, www.ilamb.org/CMIP5v6/historical]

Researchers inspect an eddy covariance flux tower at Billy Frank Jr. Nisqually National Wildlife Refuge. [Courtesy AmeriFlux]

7 | Enhancing Data and Model Exchange

The need to exchange data and models more broadly and equitably is not new but increasing application of data- and compute-intensive artificial intelligence (AI) makes it both more pressing and challenging. Participants at the AI4CH_4 workshop discussed increasing the availability and usability of data and models for a diverse set of researchers. Specifically, they believed a mechanism is needed to support data and model sharing, discovery, and reuse across three communities: researchers outside of existing networks, researchers across scientific domains, and experimentalists and modelers. The reasons behind the current lack of connections across these communities differ, but the approaches proposed in the workshop to resolve these challenges are similar in terms of incentives and infrastructure needs.

Global Problems Require Global Communities

Like many research challenges across interdisciplinary ecology and biogeochemistry fields, the availability of bottom-up observational methane data, and thus model representation, is heavily biased toward North America and western Europe (Delwiche et al. 2021). A lack of methane data in many other parts of the world prevents the research community from drawing conclusions about globally consequential issues like the global methane budget. This data gap also represents a significant deficit in diversity, equity, inclusion, and accessibility in science (Dwivedi et al. 2022).

The scientific community often must draw global conclusions based on incomplete databases wherein valuable data and mechanistic insights from poorly represented regions are missing. For example, a spatial representativeness analysis of the FLUXNET-CH4 database suggests a lack of bottom-up methane data from the humid tropics, where wetland methane emissions could be very high (Delwiche et al. 2021; McNicol et al. 2023), suggesting that we lack a truly global database and bottom-up estimate for wetland methane. AI approaches can help mitigate such data gaps and potential biases that might result, especially with continued exploration of improved gap-filling and upscaling models to estimate methane in undersampled areas (Irvin et al. 2021; McNicol et al. 2023; see Chu et al. white paper, p. 77; Feng et al. white paper, p. 106; Ricciuto et al. white paper, p. 95; Sihi white paper, p. 104). However, these issues ultimately require developing the data exchange community.

> *A lack of methane data in many other parts of the world prevents the research community from drawing conclusions about globally consequential issues like the global methane budget.*

An urgent community development need for exchanging data is deliberate and meaningful inclusion of, and capacity-building for, researchers outside North America and western Europe (Adame 2021). One example approach is the Integrated Coordinated Open Networked (ICON) science principles (Goldman et al. 2022). The ICON approach, supported by the ICON Science Cooperative (pnnl.gov/projects/icon-science) aims to use interdisciplinary integration, coordinated methods, an open research lifecycle, and broad engagement to develop science that is transferable across systems and mutually beneficial with a range of interested parties. Another example is groups working on the “Western data-bias problem” in other biogeosciences fields have identified solutions employing a people-centric approach focused on training and workshops, infrastructure and capacity building, top-down incentives (e.g., funding-based) for data sharing, and data-sharing support (Dwivedi et al. 2022; Todd-Brown et al. 2022). Under this approach, methane data from the humid tropics may exist, but local researchers may not have the time, resources, or incentives to share these data with international networks. Thus, in underrepresented regions, methane data collection and modeling could be prioritized and efforts to build communities through a people-centric and capacity-building approach should be supported.

Supporting Underrepresented Researchers Locally

The lack of time, resources, or incentives for data sharing experienced by members of the global community are also felt locally and oftentimes more acutely by underrepresented and underfunded U.S. researchers. Like many members of the global community, these researchers are also more likely to be missing from established networks. Available training; workshops; data-sharing support; explicit data management plan requirements, such as those in BER funding opportunity announcements; and accountability regarding findable, accessible, interoperable, and reusable (FAIR) data principles can promote uniform adoption of metadata standards and improve data accessibility by and for all users. These efforts would be further aided by funding to enable underrepresented researchers to curate and publish data in accordance with FAIR principles. However, given the limited breadth of U.S. funding agencies, these efforts are unlikely to achieve extensive global impact.

A more universal approach to improving access to quality data and metadata that meets community standards is to incorporate data curation steps into the research process itself. This could be achieved through a community resource that provides access to advanced analysis and modeling tools, as well as the compute power to run them. Training opportunities and outreach at all stages of development are essential to promoting broad use and buy-in. This approach provides immediate and meaningful enhancements to the productivity of individual research efforts and collectively increases the availability of quality data to the community.

BER's investment in the National Microbiome Data Collaborative (NMDC) infrastructure, specifically the Champions Program within NMDC, exemplifies this sort of community engagement, training, and support. NMDC Champions are members of the microbiome research community, some of whom focus on the methane cycle (e.g., methane production and oxidation pathways) and other carbon, nutrient, and oxidation/reduction-related pathways that interact with and affect methane flux (see Ch. 2: State of the Science, p. 9). This program provides a means of community feedback on NMDC data curation, data management, and other core activities. Furthermore, Champions promote NMDC and share content and opportunities with collaborators and the broader research community.

Bridging Research Across Domains

Understanding the methane cycle requires incorporating cross-domain knowledge. This includes understanding the physiology of microbes that have roles in the methane cycle and the physicochemical environmental conditions that influence their distribution, activity, and, ultimately, their impacts (see Ch. 2: State of the Science, p. 9). For example, model sensitivity analysis has shown the importance of capturing microbial biology in ecosystem-scale models (Song et al. 2020; Ricciuto et al. 2021; Sihi et al. 2021) and that biological understanding of methanogens and methanotrophs can be advanced by understanding the environmental context in which microbes operate. Similarly, plants play an important role in the transport and release of methane. Understanding plant physiology (especially the role of aerenchyma) and interactions with the environment is critical for predictive modeling, which can provide insight on how traits may influence plant distribution and productivity.

Cross-domain research faces many of the same communication and professional network challenges that confront researchers outside existing networks. Indeed, awareness of available data, analysis tools, workflows, and models is prerequisite to their findability and accessibility. This awareness becomes particularly challenging when research topics span traditional disciplinary and current funding boundaries; as professional networks become sparse, so too does knowledge of existing resources.

Researchers working with data outside their domain (i.e., non-domain experts) also face challenges in how to process data and verify its quality. One example is the use of genomic data in larger-scale models. Although genomic information is readily available through multiple biologically focused repositories, environmental researchers may struggle to find and access these data. They may also face challenges in interpreting quality information within sequencing data files or navigating the bioinformatic workflows needed to process sequencing data into information useful for larger-scale studies (e.g., community composition or functional potential). These researchers could benefit from better availability of data products (e.g., community or functional abundance tables) and from interactions with interdisciplinary scientists who offer different backgrounds, understanding, and expertise. Additional metadata documentation, including standardized metadata that facilitates an assessment of whether datasets can be combined, could help guide cross-domain researchers to develop improved or focused understanding of the data themselves, and ultimately increase data reusability.

Facilitating Experimentalist–Modeler Exchange

The model-experiment (ModEx) approach adopted by BER's Earth and Environmental Systems Sciences Division provides a framework that facilitates interactions between experimentalists and modelers. Regular exposure and interaction among researchers with different epistemologies helps build trust and leverage their complementary approaches to knowledge development. ModEx provides modelers with a voice on data generation by experimentalists and experimentalists with input on model structures necessary to capture understanding of the modeled system. However, forging new relationships and learning to communicate takes time, and the small overlap in professional networks may result in non-optimal pairing of models with research questions.

The challenges of exchange between experimentalists and modelers are heightened when overlaid with differences in scientific domain. The capacity to model connections between micro- and larger-scale process rates (see Ch. 2: State of the Science, p. 9) is a scientific challenge affected by the ability of modelers to find and use quality data in their model of interest. A particular challenge highlighted in this workshop and related white papers is incorporating microbiological data into larger-scale models (Xu et al. 2015; Sihi et al. 2021). In this case, models from the environmental domain require data collected from experiments and observations in the biological domain. This example requires exchange of information across two sets of language and epistemological barriers (experimentalist-modeler and biological-environmental) while also crossing traditional funding boundaries.

A consistent need of modelers (also highlighted at the AI4ESP workshop) is accessible and model-ready data (U.S. DOE 2022). This need is magnified in data-driven models that have high data input requirements and are more strongly affected by data quality. A platform that provides model-ready data could incentivize modelers to contribute modeling tools, which can likewise incentivize experimentalists to provide their data in a standardized, model-ready format. A platform providing searchable models and model-ready data also broadens the options available to both modelers and experimentalists by effectively enlarging the network of researchers with which they interact, leading to more optimal pairing of research questions with models.

Envisioning a Community Platform

A community platform for methane cycle research, available to and informed by all, that provides data, analysis, modeling tools, and compute resources can help address challenges by creating a globally accessible incentive for contributing high-quality data and modeling and analysis tools. This platform would advance the ModEx concept (see "Adapting the ModEx Framework to AI Models" sidebar, p. 4) by facilitating interactions among experimentalists and modelers and uniting a broader community through common goals. Trust could be developed through user interfaces that enable experimentalists with less modeling exposure to directly interact with and explore stable models and model components. Trust could also grow by implementing automated tracking and attribution of data and analytical tools used across the platform. The platform could serve as a community resource that incorporates a wider range of data and delivers it to a broader community. It could thus expand the advantageous impact of ModEx, perhaps introducing the concept into domains like biology where it has not yet been formalized.

Aurora supercomputer at Argonne National Laboratory. [Courtesy Argonne National Laboratory]

8 | Computing Infrastructure

Workshop discussions revealed broad consensus that artificial intelligence (AI) has great potential to advance understanding of the methane cycle, which itself represents a valuable use case for understanding AI's potential for Earth systems predictability and for strategies to realize that potential. This chapter highlights computing infrastructure investment opportunities to reach these goals, going beyond clear demands for increased processing power, data storage, and networking capacities to include software infrastructure, user-centric design, and organizational and cross-organizational processes and policies.

Fundamentally, realizing AI's potential to advance BER science and address questions regarding the methane cycle, greenhouse gasses, or any other BER-relevant topic requires a comprehensive and seamlessly interoperating system of capabilities to accelerate the model-experiment (ModEx) cycle and other AI-based capabilities. New demands on infrastructure resources will arise from AI's diverse integration of observation and simulation data, ability to facilitate data exploration and interrogation, tremendous processing power, and straightforward manipulation of large datasets.

Investing in a Comprehensive, Connective Approach

Given rapid progress in AI capabilities, infrastructure tailoring is urgently needed to realize the technology's full potential. While traditional findable, accessible, interoperable, and reusable (FAIR) data resources will continue to be centrally important, investment is needed in connective infrastructure that encapsulates multiple data resources and provides streamlined, unified access to a much larger research community. Such infrastructure would enable critical objectives detailed throughout this report. Efforts toward this comprehensive system approach to infrastructure investments should consider both advancing the ModEx framework and user-centric design and experience.

Advancing the ModEx framework

A key investment challenge is developing infrastructure that accelerates an AI-enhanced ModEx cycle and other AI-based capabilities. Many of the high-priority capabilities discussed during the workshop will require multiple iterations to prototype and then turn into useful products. Specific capabilities related to enhancing understanding of local to global methane cycling involve speeding up ModEx integration, such as by automating the collection and gap-filling of data from physical observations and experiments (see Ch. 3: Observations, Experiments, and Discovery, p. 19), parameterizing models against observed data, assimilating data, and conducting model evaluation and sensitivity analyses (see Ch. 5: Multiscale Modeling, p. 33 and Ch. 6: Data–Model Integration and Benchmarking, p. 39). Projects in critical areas, such as accelerating ModEx across the microbial-ecosystem-global spectrum, upscaling from fine-scale to large-scale environmental data, or building workflows with resources that cross agencies and national boundaries will likely need additional iterations of design and prototyping to be successful. Even for purely modeling-based studies, many current ModEx loops are manual and slow, requiring prohibitive expenditures in labor, computation, and data transfer and engineering (see "Adapting the ModEx Framework to AI Models" sidebar, p. 4).

Additional challenges associated with accelerating the development and deployment of AI-enhanced capabilities are those analogous to the challenges of realizing FAIR data (Wilkinson et al. 2016; go-fair.org/fair-principles). The value of FAIR data is limited not only by the cumulative effect of barriers to FAIRness but also by the weakest link in any of the four FAIR measures. Every barrier to a FAIR measure decreases the overall value of a given dataset, but the value drops to zero if just one of the four measures fails to be at least partially achieved.

In the case of AI, critical needs analogous to the FAIR measures go beyond data and metadata FAIRness to include compute resources, simulation capabilities, AI/machine learning (ML) capabilities, and workflow capabilities. AI adds new terminology like data hygiene and cleaning, which are applicable to FAIR data principles and raise the visibility of data quality as a dimension of reusability. Although BER's data resources (i.e., infrastructure for FAIR data) are distinct from the above needs for AI, BER's progress in resolving FAIR data challenges demonstrates its capability to meet analogously complex AI challenges.

Emphasizing User-Centric Design and Experience

The focus on accelerating the ModEx cycle and other AI-based capabilities highlights an important theme across the workshop discussions: the computational and data infrastructure required to implement AI-based solutions to address methane cycle questions must be easy to use and modify. Practically speaking, this means that different infrastructure elements must be tailored to specific domains, data, and computer scientists. Thus, user-centric design, design thinking, and user experience/user interface (UX/UI) expertise could be high-impact opportunities if included in all infrastructure investments throughout project lifecycles, including planning, prototyping, testing, community engagement, and training and support.

Maximizing Existing Technologies

Existing infrastructure technologies have the potential to meet the needs and opportunities outlined in this report. They include: (1) compute and networking resources, (2) FAIR data resources, and (3) scientific workflows.

Compute and Networking Resources

Many potentially high-impact AI applications discussed during the workshop involve tying together currently disconnected datasets; harmonizing, completing, and otherwise transforming data; and connecting data and models. Infrastructure roles include software and systems that facilitate these tasks but are not specific to a particular dataset or type.

Subject-specific, non-infrastructure roles are discussed in other sections of this report. Diverse aspects of advancing AI to study the methane cycle necessitate an infrastructure that can support large amounts of data at many scales and modes of bioscience and Earth science data. Among the opportunities and challenges to realizing such infrastructure are the following:

- Training AI models, especially foundation models, requires vast amounts of computing resources.
- Traditional data assimilation models (e.g., ensemble Kalman filters) are computationally expensive.
- Many, if not most, processes in the methane cycle possess aleatoric and epistemic uncertainties. This situation requires sizable AI model ensembles to quantify uncertainties or analyze sensitivities. In addition, AI model ensembles require substantially greater compute resources than non-ensemble approaches, but they enable thorough calibration of and trust in AI-based and AI-enhanced models.
- Training AI models requires increased data reusability and, in turn, computing capacity. Such increases will continue to accelerate as communities deploy large language models and generative AI to explore and curate data and metadata. Such deployments will make data, models, and workflows more FAIR to non-domain experts.
- As standardized workflows enhance nonspecialist access to existing knowledge, data, and models, demand for computing resources will rise correspondingly.
- Access to significant computing resources can dramatically facilitate use of proprietary data, presenting both challenges and opportunities. For example, to address privacy concerns, BER could establish agreements permitting learning on encrypted data, although it is more computationally expensive than learning on unencrypted data.
- The availability of adequate computing resources can accelerate scientific discovery by enabling model-guided, small-scale experiments. However, domain science researchers must be free to use the resources without having to manage capacity constraints (e.g., compute, storage, or network) via standardized workflows.
- Anticipated improvements in simulation-based models will further increase computing needs and costs. For example, multiscale modeling advances are needed, such as incorporating (1) microbial models into larger-scale (and expensive) environmental models and (2) uncertainty quantification methods into both fine-scale and large-scale models.
- Advanced Scientific Computing Research (ASCR) is deploying exascale systems at its high-performance computing (HPC) centers: Argonne Leadership Computing Facility, Oak Ridge Leadership Computing Facility, and National Energy Research Scientific Computing Center. Collectively, the capabilities at these facilities represent a vast compute resource equipped with state-of-the-art AI hardware, but access to the facilities is unevenly applied in terms of users, scale of access (node or core hours), and annual or semiannual peer review processes. Similarly, different BER user facilities, such as the Environmental Molecular Sciences Laboratory and Atmospheric Radiation Measurement (ARM) user facility, manage their own computing resources and data environments, which also have different modes of access.

To realize AI's potential, network and storage resources and their management are essential considerations during strategy development and planning. First, network bandwidth can be overwhelmed by exponential growth in the size of data streams from field sensors, laboratory instruments, observational imaging, and simulations (e.g., large ensembles of climate simulation results or regional-scale simulations under multiple scenarios). Network and storage investments will yield maximum benefit if designed for long-term scalability based on current trends. Second, improving data FAIRness and data connectivity, as well as democratizing standardized workflows, will increase data transfer needs substantially. These increased needs are particularly important for enabling and encouraging the use of more datasets in model training efforts across more institutions. Third, drones and other autonomous or semi-autonomous sensor platforms (e.g., for measuring methane fluxes) provide additional challenges for the communication infrastructure.

As discussed in the Emerging Technology Development section (see p. 55), emerging technologies such as edge computing, low-power computing, and 5G networks offer important opportunities to mitigate associated burdens on networking and compute needs. More generally, ASCR's Energy Sciences Network (ESnet) and planned upgrades provide world-class capacity among key research sites around the world, specifically for DOE national laboratories and scientific user facilities. ESnet also offers excellent pairing connectivity to the networks and compute resources of public cloud service providers. Furthermore, ESnet ambassadors, located at each national laboratory and user facility, ensure high-quality responsiveness to challenges, issues, and needs. Finally, as discussed in the Scientific Workflows section (see p. 53), network infrastructure burdens due to these factors can be significantly alleviated by deploying workflow management capabilities that minimize data transfer, moving compute capabilities to the data rather than the traditional approach of moving data to the compute capabilities.

FAIR Data Resources

To accelerate ModEx cycles, researchers require instantaneous access to scalable computing resources. The value of these resources is maximized when access is seamless in both effort and start-up time and when scale-up barriers are minimized. BER has already made substantial and reasonably comprehensive investments in FAIR data resources, including the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), Earth System Grid Federation (ESGF), National Microbiome Data Collaborative (NMDC), DataONE network, and National Virtual Climate Laboratory.

Reflecting the success of these investments, workshop discussions focused on the positive impacts of increasing the usability and reusability of existing data. Such investments could, for instance, emphasize building additional infrastructure around these existing resources to facilitate connectivity; advance FAIR principles, especially for nonexpert users; and expand the number of data resources available to the community. In an idealized future, web portals could employ a high-quality user-centric design that enables self-service access to relevant data from multiple resources, including AI-enhanced recommendation capabilities.

> *To accelerate ModEx cycles, researchers require instantaneous access to scalable computing resources.*

An emerging complement to FAIR data infrastructure is providing compute resources alongside data. This could mean partnering with large computing resources to decrease data transfer (or to "bring the compute to the data"). This approach is especially important to consider given the: (1) large amounts of data needed to train a single AI model, (2) envisioned growth in the number of AI models that scientists train, and (3) increased data reusability in training AI models. Such infrastructure, which may be partially supplied by workflow capabilities (see Scientific Workflows, p. 53), improves the availability of data and models to diverse users and accelerates knowledge transfer.

Many studies of the methane cycle need, or would benefit from, seamless access to data distributed across multiple resources. Given the diversity of data sources, integration across institutions to support a federated data access policy would enable each data resource owner to maintain key autonomy while benefiting from the collective. Federated structures could also facilitate agreements that enable the community to leverage proprietary data; emerging technologies for learning on encrypted data may lower barriers even further (see Emerging Technology Development, p. 55). Common practice has been to build unified data portals, but the increasing complexity of the data landscape is driving a shift toward interoperating resources powered by a distributed metadata search and application programming interfaces (APIs) for data access. Such investments could then enable AI-aided approaches to data harmonization.

The need for standardized metadata surfaced repeatedly in workshop discussions, especially in the context of genomics data collected for the study of methane cycling, and represents an essential component of the consensus view that investments must enhance capabilities that leverage existing knowledge. Tools to manage and connect metadata schema could therefore be invaluable, especially given that many DOE mission–science questions involve mechanisms and processes that bridge gaps in organizational and domain science knowledge (e.g., among roots, the rhizosphere, microbial communities, soil structure, and other physical and geochemical data).

Although different research subfields may have metadata schema for individual topics (e.g., rhizosphere microbes), barriers to connecting these datasets persist. Innovations in methods and deployment of software that facilitates prototyping and testing different connections between multiple metadata schemas would be valuable. Also potentially useful are systems that distribute metadata across partnering sites. Gaps exist in researchers' ability to perform comprehensive, multisite searches without needing to master different data resources. A common software platform equipped with access APIs could be useful in this regard. Other computing capabilities, including simulation codes, trained AI/ML models, and computational workflows, could be supported with FAIR principles akin to observational data or simulation output data (Wilkinson et al. 2016; Goble et al. 2020; see Scientific Workflows, this page).

Scientific Workflows

Scientific workflow capabilities represent a core infrastructure element supporting AI's potential for understanding the methane cycle. Workflows play a central role in bridging computational, experimental, and observational research and represent a critical enabling technology for embedding AI throughout BER research. Cutting-edge data-driven studies are expected to increasingly require resource integration along a computing continuum, from edge to local (i.e., on premise), cloud resources, and ASCR HPC centers. Currently, tools and systems that enable easy leveraging of fully autonomous or self-driving sensing and experimental systems are lacking.

Workflow capabilities also lower barriers to interdisciplinary collaboration on grand challenges and offer solutions to research bottlenecks, such as bidirectional linking between biological data and environmental models. As such, workflows are a key component of AI infrastructure. Overall, the community needs better tools to facilitate the exploration, development, refinement, and distribution of such solutions.

The impacts of AI workflow infrastructure are comparable in significance to the impacts of infrastructures that democratized data and HPC. Public data resources supporting FAIR principles accelerated research by enabling researchers anywhere to access large, diverse datasets without personal or professional connections. In HPC, standards for parallel programming accelerated research by enabling scientists to share and compare large-scale, high-performance algorithms and computations using different HPC resources. This advance freed computational scientists from having to re-implement each algorithm on their parallel hardware. In the era of AI and AI-enhanced ModEx, scientific productivity would be enhanced if workflow development, dissemination, and sustainability were similarly facilitated, freeing scientists from the tedious and error-prone manual labor of connecting available data and compute with the tools they need.

Supporting Workflows

In the past decade, the research community has made substantial progress on computational, communication, ownership, and provenance challenges facing development and support of scientific workflows (e.g., da Silva et al. 2021, Pegasus, PARSL, funcX). Other core technologies facilitating workflows include low-code and no-code environments and standardizing containers (e.g., Docker, Singularity/Apptainer, Shifter for HPC, and Advanced Terrestrial Simulator).

Workshop discussions highlighted applications with important workflow uses:

- Standardizing methane observation data pipelines using workflow platforms improves quality control, which is important for upscaling (Hu et al. 2022).
- Combining chamber and tower methane flux data increases spatial data points for global flux upscaling, which is valuable due to the disparity in the number of chamber flux sites compared to towers.
- Developing workflows to reduce barriers to creating data products can include, for example, capabilities for automating inclusion of new data; testing different approaches to resolve inconsistencies in data reporting; providing flexible, on-demand gap-filling; conducting large-scale sensitivity analyses in complex systems; and using high-resolution surrogate models for upscaling.
- Examples of scientific workflows that integrate uncertainty quantification with bridging across scales include nowcasts of methane emissions featuring uncertainty arising from area fluctuations, and high-resolution regional- to global-scale maps of methane fluxes and uncertainties under historical (i.e., reanalysis) and future scenarios.
- Developing better workflow capabilities can improve coordination and execution between sensing devices and computing, yielding new approaches to data collection and advancing domain science by optimally informing underlying models (e.g., supporting active learning to identify optimal measurement locations).
- Creating reproducible workflows greatly facilitates assessments related to policies, such as reservoir management (e.g., effects of drawdown timing) and the current and future roles of croplands in methane production and consumption. Policy and semiautonomous data-collection applications provide an additional motivation for investing in robust workflow capabilities. Namely, they are important for understanding which algorithms and execution methods meet particular needs, such as response time or resiliency.
- Workflows enhance access to verification and model updating because they facilitate repetition of model training on complete and distributed datasets.
- Developing trust in AI requires running various models in sync, which is more easily performed with standardized workflows. A comprehensive workflow ecosystem also helps standardize procedures and best practices on data calibration and quality control/quality assurance that may be performed at the edge or offline, while critically preserving flexibility for site-specific or sponsor-specific needs.
- Building pipelines in which new measurements, especially time-resolved measurements, can automatically trigger model updates are important workflow features (e.g., waterbody area measurements and methane uncertainty models or data assimilation workflows that use remote-sensing data pipelines as a basis for Bayesian sensitivity analysis).
- Lowering barriers to productive AI application can be achieved using workflows, whether for model-data integration, ModEx, or other purposes, especially in conjunction with easy mechanisms for searching and using data across multiple sites and agencies. A specific computing example involves opportunities afforded by relatively well-developed areas of AI technology, such as 2D and 3D image analysis and processing.

Workflow Impacts

With respect to data harmonization, improved workflow tools can increase data reusability for training AI models, enhance capabilities to leverage existing knowledge in any digital form, facilitate AI-aided approaches to enhance data harmonization, and support metadata standardization. In combination with new data collection technologies, distributed workflows facilitate flexible, high-resolution sampling as well as high-throughput laboratory-scale data generation for model training. Workflows leveraging distributed sensor networks, autonomous platforms, and edge computing (see Emerging Technology Development, p. 55) may benefit from additional specific infrastructure.

An important accessibility gap impacted by workflows will be the transfer of insights through models and data, especially across the biological and environmental sciences which are both needed to understand the methane cycle. This need is especially clear for multiscale modeling. Workflow capabilities accelerate the research design-build-test-learn (DBTL) cycle for building new modeling capabilities because they facilitate rapid testing of ideas through prototyping. Workshop examples included approaches for improving model performance through data assimilation, identifying and resolving key parameters and structural uncertainties, developing and evaluating surrogate models, integrating surrogate models with physics-based models, and incorporating human dimensions into numerical and AI models.

Finally, the impacts of workflow infrastructure on community development should not be underestimated. Workshop participants highlighted needs for making data and models available to diverse sets of users; easily providing compute alongside data; and developing a multidisciplinary workforce with shared baseline fluency in distributed data, metadata tools, and workflow development and use.

Emerging Technology Development

This section highlights opportunities for infrastructure research where the feasibility of delivering production-grade systems is not yet well understood. Examples include developing automated or semi-automated infrastructure for sensing, edge computing, and wireless technologies which may be of interest to DOE's ASCR program or to BER and ASCR jointly.

AI promises synergies with emerging approaches, including edge computing and multiple new sensing systems, and offers critical new opportunities to address data gaps and accelerate model improvements. Compared to delivering a high-performance AI ecosystem using current technologies, integrating emerging technologies in high-impact applications presents distinct challenges. However, these challenges can be mitigated through investments that simultaneously advance the capability and understanding of how to embed the capability into practice. This strategy is especially important given AI's rapid progress.

> *AI promises synergies with emerging approaches, including edge computing and multiple new sensing systems, and offers critical new opportunities to address data gaps and accelerate model improvements.*

Computer science investments will be important for addressing knowledge gaps in the integration of experiments with HPC and cloud computing with scientific workflow resilience. Similarly, but at a smaller scale, computer science and mathematics research can advance edge computing capabilities for automating data collection and inference. For instance, depending on the scale of computing capability in edge devices, the connected sensing devices may be able to switch between passive sensing, active, and smart modes (e.g., autonomous responses to anomaly detection). In this space, DOE investments in co-design create opportunities for powerful edge computing capabilities that deliver high-fidelity AI models while using very little power or network bandwidth.

One area of potential shared interest between ASCR and BER involves rapidly prototyping and interrogating modified schema or algorithms that harmonize metadata. Approaches are needed given the importance of standardizing metadata under uncertainty (i.e., evolving scientific understanding), the possibilities for AI to accelerate metadata harmonization and completion, and the increased prevalence of large, distributed datasets. Another needed approach involves efficiently exploring distributed data prior to use in compute-intensive workflows.

To address challenges associated with integrating AI with data federations across institutional boundaries, ASCR and other agencies have already begun investing in mathematics and algorithms to address longstanding challenges in accessing proprietary data. Examples include federated learning (i.e., learning in a distributed sense) and learning on encrypted data. Cyberphysical systems and workflows represent another important opportunity. Workshop discussions highlighted the potential of autonomous laboratory and field sensor systems to address critical data needs, such as through flexible but high-resolution sampling. Shared interests in applied mathematics and computational needs for multiscale models are discussed significantly in Ch. 5: Multiscale Modeling (see Advancing Predictive Capabilities, p. 35) and represent an area of ongoing shared investments through, for example, the DOE Scientific Discovery through Advanced Computing (SciDAC) Partnerships program.

Key opportunities in emerging computing and networking technologies include:

- Field-based sensors that leverage next-generation 5G networks and are equipped with pre-trained AI models to identify anomalies or change their behavior according to predefined rules or simulation-based model predictions.
- Extreme-scale storage of large ensembles of climate simulation results or regional-scale simulations under multiple scenarios, which include specific outputs related to the methane cycle.
- Edge computing to enhance experimental and sampling design at scales ranging from laboratory to field.
- Workforce development is an important space for shared investments for early-career researchers from diverse backgrounds to work fluently within an integrated infrastructure.

Appendix A

Agenda

Artificial Intelligence for the Methane Cycle (AI4CH_4) Workshop Series

March 3

12:00–12:20 p.m.	Workshop introduction and charge
12:20–1:50 p.m.	Brief presentations highlighting gaps and opportunities
1:50–2:15 p.m.	Group Q&A
2:15–2:45 p.m.	Break
2:45–4:40 p.m.	Breakouts for participant introductions and ideation
4:40–5:00 p.m.	Report back to full group

March 10

12:00–2:15 p.m.	"Improving Predictions from Fundamental Microbiology" Discussion Section
2:15–2:45 p.m.	Break
2:45–5:00 p.m.	"Environmental Controls and Empirical Relationships" Discussion Section

March 17

12:00–2:15 p.m.	"Targeting Field Measurements and Observations" Discussion Section
2:15–2:45 p.m.	Break
2:45–5:00 p.m.	"Data-Model Integration" Discussion Section

March 24

12:00–2:15 p.m.	"Multiscale Modeling" Discussion Section
2:15–2:45 p.m.	Break
2:45–5:00 p.m.	Breakout by session topic for workshop synthesis and report drafting

Topics for Each Discussion Session

- Knowledge gaps and scientific questions
- Characteristics and challenges of specific data and models
- Related algorithms, infrastructure, and their potential to address gaps
- New observations, measurements, and experimental design needed
- Data products (QC, UQ, harmonization, benchmarks) needed

Appendix B

Workshop Participants

Chair

Pamela Weisenhorn, *Argonne National Laboratory*

Co-Chairs

James Ang, *Pacific Northwest National Laboratory*

Jaydeep Bardhan, *Pacific Northwest National Laboratory*

Maxwell Grover, *Argonne National Laboratory*

Forrest Hoffman, *Oak Ridge National Laboratory*

Daniel Ricciuto, *Oak Ridge National Laboratory*

Charuleka Varadharajan, *Lawrence Berkeley National Laboratory*

Organizer

Olga Tweedy, *U.S. Department of Energy, former AAAS fellow*

Attendees

Dionysios Antonopoulos, *Argonne National Laboratory*

Sebastien Biraud, *Lawrence Berkeley National Laboratory*

Kristin Boye, *SLAC National Accelerator Laboratory*

Clifton Bueno de Mesquita, *DOE Joint Genome Institute*

Min Chen, *University of Wisconsin–Madison*

Housen Chu, *Lawrence Berkeley National Laboratory*

Scott Collis, *Argonne National Laboratory*

Ewa Deelman, *Information Sciences Institute*

Kyle Delwiche, *University of California–Berkeley*

David Durden, *National Ecological Observatory Network*

Sha Feng, *Pacific Northwest National Laboratory*

Etienne Fluet-Chouinard, *Pacific Northwest National Laboratory*

Behzad Ghanbarian, *Kansas State University*

Dalei Hao, *Pacific Northwest National Laboratory*

Todd Hay, *Pacific Northwest National Laboratory*

Christopher Henry, *Argonne National Laboratory*

Alison Hoyt, *Stanford University*

Satish Karra, *Environmental Molecular Sciences Laboratory*

Sara Knox, *University of British Columbia*

Daniel Krofcheck, *Sandia National Laboratories*

Fa Li, *Sandia National Laboratories*

Johnny (Liujun) Li, *University of Idaho*

Licheng Liu, *University of Minnesota*

Tiia Määttä, *University of Zürich*

Avni Malhotra, *Pacific Northwest National Laboratory*

Sparkle Malone, *Yale University*

Melanie Mayes, *Oak Ridge National Laboratory*

Jorge Mazza Rodrigues, *University of California–Davis*

Gavin McNicol, *University of Illinois–Chicago*

Christof Meile, *University of Georgia*

Kendalynn Morris, *Pacific Northwest National Laboratory*

Maruti Mudunuru, *Pacific Northwest National Laboratory*

Robinson Negron-Juarez, *Lawrence Berkeley National Laboratory*

Vincent Noel, *SLAC National Accelerator Laboratory*

Michael Nole, *Sandia National Laboratories*

Genevieve Noyce, *Smithsonian Environmental Research Center*

Youmi Oh, *NOAA Global Monitoring Laboratory*

Edward O'Loughlin, *Argonne National Laboratory*

Benjamin Poulter, *National Aeronautics and Space Administration*

Peter Regier, *Pacific Northwest National Laboratory*

William Riley, *Lawrence Berkeley National Laboratory*

Rajesh Sankaran, *Argonne National Laboratory*

Debjani Sihi, *Emory University*

Hyun-Seob Song, *University of Nebraska–Lincoln*

Yang Song, *University of Arizona*

Jemma Stachelek, *Los Alamos National Laboratory*

Benjamin Sulman, *Oak Ridge National Laboratory*

Nathan Tallent, *Pacific Northwest National Laboratory*

Zeli Tan, *Pacific Northwest National Laboratory*

Jinyun Tang, *Lawrence Berkeley National Laboratory*

Hoang Tran, *Pacific Northwest National Laboratory*

Susannah Tringe, *Lawrence Berkeley National Laboratory*

Jiaze Wang, *University of Maine*

Nicholas Ward, *Pacific Northwest National Laboratory*

Kenneth Williams, *Lawrence Berkeley National Laboratory*

Xiaofeng Xu, *San Diego State University*

Zutao Yang, *Stanford University*

Fenghui Yuan, *University of Minnesota*

Kunxiaojia Yuan, *Lawrence Berkeley National Laboratory*

Qing Zhu, *Lawrence Berkeley National Laboratory*

Observers

Dawn Adin, *U.S. Department of Energy*

Jennifer Arrigo, *U.S. Department of Energy*

Xujing Davis, *U.S. Department of Energy*

Gerald Geernaert, *U.S. Department of Energy*

Kim Hixson, *U.S. Department of Energy*

Justin Hnilo, *U.S. Department of Energy*

Renu Joseph, *U.S. Department of Energy*

Resham Kulkarni, *U.S. Department of Energy*

Shing Kwok, *U.S. Department of Energy*

Sally McFarlane, *U.S. Department of Energy*

Vijay Sharma, *U.S. Department of Energy*

Daniel Stover, *U.S. Department of Energy*

Amy Swain, *U.S. Department of Energy*

Boris Wawrik, *U.S. Department of Energy*

Tristram West, *U.S. Department of Energy*

Appendix C

White Papers

U.S. Department of Energy, Office of Science, Biological and Environmental Research Program

Artificial Intelligence for the Methane Cycle (AI4CH_4) Workshop Series

Purpose

A joint Biological Systems Science Division and Earth and Environmental Systems Sciences Division workshop series will be held to advance predictive understanding of the methane cycle, with explicit consideration of the role of advanced statistical approaches [including artificial intelligence (AI) and machine learning (ML)], federated data resources, smart sensors, and distributed continuum computing. This workshop will bring together a diverse group of researchers across career stages, including members of the methane modeling and measurement/observation communities and subject matter experts from the larger computational ML, hardware/software co-design, edge systems, and networking/communications communities. A particular emphasis of this workshop is bridging the gap in research knowledge across scales, utilizing new approaches across scales with an emphasis on biological and environmental scientific domains. The purpose of this announcement is to solicit white papers from the scientific community that focus on the methane cycle and/or development and application of advanced computational methodologies, including AI and ML. White papers will be used to guide the workshop planning and invitations.

Structure of White Papers

White papers should be prepared using the following outline and may be up to a maximum of 2 pages long (12-point font, not including the optional References sections).

1. Title
2. Authors/Affiliations: List in order of largest contribution
3. Focal Area(s): One or two sentences only; see last paragraph for list
4. Science or Technological Challenge: Short statement describing the area addressed by the white paper
5. Rationale: Description of the research needs/gaps, the barriers to progress, and the justification for and benefits associated with the proposed approach
6. Narrative: Brief scientific and technical description of the scientific objective or approach; activities that will advance the science; and specific field, laboratory, model, synthesis, and/or analysis examples
7. References (Optional)

Authors are limited to one submission as lead author but may participate as a co-author in other submissions. Teaming is encouraged to reduce the reviewing workload. Multi-institutional responses are welcome; however, a clear lead who can speak authoritatively on the white paper contents should be identified. [Note: Protected information should not be included in white papers, but instead should be shared directly with the appropriate U.S. Department of Energy (DOE) program manager(s).]

Submission Process

White papers must be submitted as PDF files by 5:00 p.m. EST on 17 February, 2023, using this Google Form. After the completion of the workshop, white papers will be posted publicly through the AI4ESP workshop website (www.ai4esp.org/related).

Background

A particular interest within the DOE Office of Science's Biological and Environmental Research (BER) program is the carbon cycle. This cycle includes the stocks and fluxes of carbon in its various forms throughout the biosphere, as well as its impact on both biological systems behavior and climate. Methane is an important component of the carbon cycle, with 20-30 times the radiative forcing of carbon dioxide. Methane enters the atmosphere via several natural and anthropogenic sources, and methane cycling has been of intense interest to the Earth science community. Despite this, there are large uncertainties in land-atmosphere exchange estimates of methane in global models, due in part to high spatial and temporal variability of these fluxes and related processes. This variability has many causes, including: environmental sensitivity of the microorganisms, the importance of interactions among species for both production and consumption processes, the importance of abiotic processes in mediating release, and biotic and abiotic barriers to flow. Additionally, these uncertainties are affected by the sparsity of process-relevant environmental data, differences in measurement approaches and frequencies across a wide range of scales, and uncertainty in the measurements themselves.

The recent BER-ASCR Artificial Intelligence for Earth System Predictability (AI4ESP) virtual workshop series, held during October and December 2021, identified next-generation capabilities "to more radically and aggressively advance prediction capabilities in the climate, Earth, and environmental sciences through the use of modern data analytics and artificial intelligence." Many of these approaches may be applicable to the data and modeling challenges relevant for developing a predictive understanding of the methane cycle across scales with potential contributions to predictive understanding of both the biological and environmental components of the methane cycle. For example, deep learning algorithms, surrogate models, and multi-fidelity hybrid (ML and process) models have the potential to address challenges with both scaling and heterogeneity of microbial processes. AI-enabled technologies can be used to obtain automated measurements of methane flux and process rates to better capture the high spatiotemporal variability of methane release and flux and process response to lab and field manipulations, including those mimicking extreme events. Causal inference and information theory coupled with AI approaches, may help enable a deeper understanding of microbial and biogeochemical drivers of the methane cycle. To achieve these gains, there is a need to have accessible and synthesized datasets capturing various aspects of the methane cycle across scales.

Call for White Papers

White papers should be framed around one or more of the following focal areas:

- Key uncertainties and knowledge gaps where new methodology, infrastructure, or technology can advance predictive understanding of the methane cycle. This advance can be realized within a scientific domain, across domains, or in model improvements.
- A solution to a key challenge in implementing AI approaches (e.g., improving uncertainty quantification, federated learning) across the biological and environmental science domains as it pertains to the methane cycle.
- The importance of high-potential datasets (e.g., genomics or other omics data, eddy covariance networks, remote sensing) or how the combination of data across spatial or temporal scales or scientific domains may lead to new scientific insights, either within or across fields. Where relevant, white papers should highlight how advanced statistical and numerical methods can be used to realize this insight. How automated or real-time data capture and processing or federated learning can be used to address issues of spatial and temporal heterogeneity and sparsity (e.g., through improvements in measurement coverage or uncertainty quantification).
- Approaches that support the transfer of knowledge gained in the laboratory to make predictions in the field and vice versa.

Partitioning Net Wetland CH_4 Emissions into Production and Oxidation Components Using Ecosystem-Scale Flux Measurements and Physically Guided Machine Learning

Qing Zhu,[1] William Riley,[1] Jinyun Tang,[1] Robinson Negron-Juarez,[1] Housen Chu,[1] Kunxiaojia Yuan,[1] Gavin McNicol,[2] Min Chen,[3] Fa Li[3]

[1]Lawrence Berkeley National Laboratory, [2]University of Illinois at Chicago, [3]University of Wisconsin–Madison

Focal Areas

- 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 CH_4 emissions result from production sources and oxidation sinks that are controlled by different microbial groups (i.e., methanogens and methanotrophs). These production and oxidation rates have different dynamics and exhibit a wide range of responses to environmental changes. The net emissions also depend on transport processes (e.g., aerencnyma, ebullition, diffusion). However, ecosystem-scale long-term observations (e.g., FLUXNET-CH4; Delwiche et al. 2021) only measure net CH_4 emissions, which hinders predictive understanding of wetland CH_4 emissions across space and time.

A robust flux partitioning algorithm is urgently needed to decompose observed net emissions into gross production and oxidation rates. Such new datasets can be used to improve predictions of future wetland CH_4 emissions as well as spatial upscaling across heterogeneous landscapes.

Rationale

Wetland CH_4 emissions represent ~20% to 30% of global CH_4 emissions, and these emissions are increasing due to ongoing climate warming because the radiative power of CH_4 is ~30 times stronger than CO_2 over a 100-year time horizon. Classic approaches to estimate wetland CH_4 emissions used either process-based bottom-up (BU) models that directly simulated wetland biogeochemical processes or top-down (TD) transport models that indirectly inferred wetland CH_4 emissions based on atmospheric CH_4 concentrations. Existing BU modeling studies showed some progress in capturing the observed CH_4 emissions at a handful of FLUXENT-CH4 sites after careful calibration. However, BU models still suffer from large parametric uncertainty and use incomplete biogeochemical theories. Furthermore, TD models often use prior information derived from BU model estimates and other non-wetland surface CH_4 emissions that inevitably introduce uncertainties. The most recent Global Carbon Project methane budget revealed ~30 $TgCH_4$ discrepancies in the magnitude, inter-annual variability, and long-term trends of BU and TD model estimates of wetland CH_4 budgets.

The ongoing synthesis efforts at FLUXNET-CH4 sites provide useful data to parameterize BU models. However, the net CH_4 emissions from the FLUXNET-CH4 dataset do not provide sufficient constraints on the CH_4 biogeochemical cycle.

Minimally, a BU model requires methane gross production and oxidation rates to constrain methanogenesis and methanotrophic processes, respectively. Thus, developing a robust partitioning algorithm for FLUXNET-CH4 CH_4 emissions becomes a critical research need to improve process understanding and model predictability of wetland CH_4 cycle.

Narrative

Our overall objective is to robustly partition observed FLUXNET-CH4 net CH_4 emissions into production sources and oxidation sinks using physically guided machine learning (PGML). We will leverage the existing FLUXNET-CH4 dataset (Delwiche et al. 2021) and new wetland sites in South America to generate global datasets of wetland CH_4 production and oxidation rates. Although the measurements of CH_4 emissions are far fewer than those of CO_2 fluxes, we expect to overcome this data limitation by developing an advanced PGML model. Unlike traditional machine learning, which depends entirely on the information content of a big dataset, PGML can combine physical principles and ecological theory to leverage the information contained in a more limited dataset to understand and predict the dynamics of target processes. Our previous work on PGML has demonstrated promising model performance at temperate and high-latitude wetland sites (Yuan et al. 2022). In our previous version of PGML, we have successfully integrated causal knowledge of how CH_4 emissions interact with physical and biological factors. Here, we will further develop the PGML model to include (1) distinct temperature sensitivities of methanogens and methanotrophs in our PGML model, (2) pretraining with synthetic data from a more mechanistic microbial model, and (3) constraints on model structure based on the knowledge of methane process interactions.

References

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–89.

Yuan, K., et al. 2022. "Causality Guided Machine Learning Model on Wetland CH4 Emissions Across Global Wetlands," *Agricultural and Forest Meteorology* **324**, 109115. DOI:10.1016/j.agrformet.2022.109115.

Upscaling Global Wetland Methane Emissions with Causality Guided Machine Learning

Kunxiaojia Yuan,[1] Qing Zhu,[1] William Riley,[1]
Gavin McNicol,[2] Fa Li,[3] Min Chen[3]

[1]Lawrence Berkeley National Laboratory, [2]University of Illinois at Chicago, [3]University of Wisconsin–Madison

Focal Areas

- 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 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 miscapture 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 areas, and wetland hot spots in the boreal Arctic area).

Therefore, a more advanced ML model that can be robustly trained by existing datasets and more *in situ* observations is urgently needed to generate a reliable global wetland methane emission dataset. Such upscaled datasets 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 CO_2 with a Global Warming Potential (GWP) 28–34 times that of CO_2 over a 100-year time horizon (IPCC 2013). Wetlands are the largest natural source of global CH_4, contributing 20% to 30% to global CH_4, 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 in benchmarking bottom-up and top-down models. However, most currently 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 multisourced 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 nonhotspot 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 in a recent global carbon project analysis, and further investigate the predominant drivers which regulate the long-term trend and variability of wetland CH_4 emissions.

References

Intergovernmental Panel on Climate Change. 2013. "Ch. 6: Carbon and Other Biogeochemical Cycles." In *Climate Change 2013 The Physical Science Basis*. Cambridge University Press, Cambridge, U.K.

Saunois, M. et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data*, **12**(3): 1561–1623.

Yuan, K., et al. 2022. "Causality Guided Machine Learning Model on Wetland CH_4 Emissions Across Global Wetlands," *Agricultural and Forest Meteorology* **324**, 109115.

Cloud and HPC Ecosystems for Scientific Experiments

Nathan Tallent, Steven Spurgeon

Pacific Northwest National Laboratory

Focal Areas

- Knowledge gaps in methodology and infrastructure for workflow execution.
- Challenges implementing AI approaches for automating feedback to scientists or instruments.

Science or Technological Challenge

Future scientific discovery requires automating data-driven feed- back to scientists or instruments to handle the full array of data generated by modern hardware, rapidly make decisions, and extrapolate beyond limits of any one experimental dataset. Examples range from automating analysis from simulations, sensors, and AI-driven models to forming real-time loops that can guide instruments or automate experiments, such as high-resolution analysis of material and chemical systems (Akers et al. 2021; Olszta et al. 2022). To achieve new levels of automation with machine reasoning, we must harness distributed scientific workflows that can exploit continuum computing ecosystems to meet both cost budgets and quality of service, i.e., response-time and resiliency.

A related need is unsupervised learning, required because of the infeasibility of labeling, which usually requires massively distributed training and substantial computational resources. The task of scientific discovery is often ill-suited to transfer learning approaches, which may lack generalizability to accurately describe or assess new experimental features.

We envision coordinated teams of domain scientists and computer scientists that design workflows to meet budget and quality of service requirements. Domain scientists would establish context by defining the domain challenges that represent fundamental limitations imposed by current computing solutions. Computer scientists leverage frameworks for co-design of cost, response time, and resiliency to guide workflow executions that combine multisystem resources, especially near-instrument computing, facility HPC, and cloud (see Fig. 1). Cloud can complement DOE computing beyond on-demand scale-out because it now drives computing trends by showcasing novel systems (quantum), new platforms (TPUs), system virtualization (containers, serverless), and machine learning frameworks (PyTorch, TensorFlow).

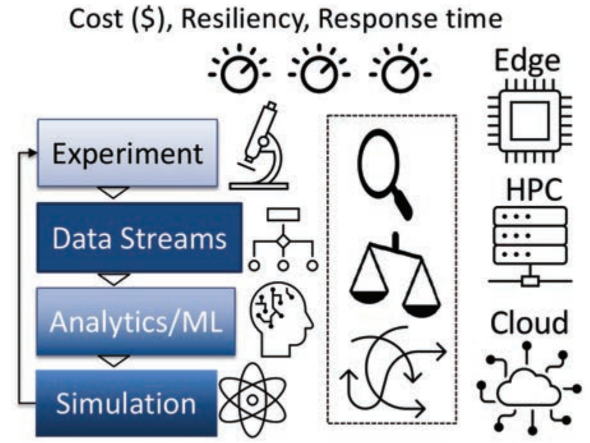


Fig. 1. To accelerate the automated data-driven feedback necessary for future scientific discovery, it is necessary to harness distributed scientific workflows that execute on continuum computing ecosystems and that meet cost budgets, time deadlines, and resiliency requirements.

Rationale

Today's tools naively execute workflows on multiple systems. Customizing data movement and resiliency is critical to meet time constraints but must be done manually—a cumbersome and error-prone process. Cloud's Function-as-a-Service (FaaS) model is attractive for cost and availability but has not been designed for meeting workflow time constraints.

Meeting cost, response-time, and resiliency raises fundamental research challenges. Costs in cloud vary widely based on service (static instance vs. stateless container), hardware, and availability-resiliency guarantees. The execution time of workflow tasks not only depends on partitioning and assignment, but on prioritizing task vs. data movement, data layouts, data movement schedules, and data caching and consistency policies. Task resiliency is usually inversely related to time. Even worse, it is usually implicit and fixed, and within workflows results in redundant recovery efforts. Further, workflows can change resiliency semantics (e.g., "final answer" tasks still need checkpointing) but "exploratory tasks" can use best-effort. Finally, workflows that vary with input and time require dynamically adaptive policies.

Vision. There is a critical need for automated co-design techniques that can address the following questions:

- Given a target budget and quality of service, what is the best selection of resources across facility, cloud, and edge (near-instrument) resources to create a virtual platform?
- What is the best assignment of policies for task placement, data movement, and task resiliency?
- Given a set of fixed resources and optional target budget, what is the range of pareto-optimal execution policies and their corresponding tradeoffs?

Narrative

Our approach is to develop transferable co-design techniques and tools within the four thrust areas below. We target workflows for rapid scientific exploration; most are input/output (I/O) intensive and coordinated with workflow managers. The co-design framework reasons about the cost-time tradeoff space for policies that meet given constraints.

Workflow-guided characterization of performance and resiliency to develop models that drive co-design. To reason about co-design tradeoffs, we develop workflow-specific models of data lifecycles and resiliency relative to key workflow parameters. We use distributed and scalable workflow introspection within I/O middleware to capture data lifecycles between workflow tasks (Suetterlein et al. 2019; Friese et al. 2020; Kilic et al. 2022).

Coordination and resource partitioning for cloud and HPC ecosystems. To meet cost and quality of service constraints in Cloud and HPC Ecosystems, we will efficiently map tasks to execution policies and multi-system resources (Suetterlein et al. 2019). The scheduler compares predictions of task performance with execution dynamics and, if necessary, adopts recommended alternatives.

Optimizing I/O middleware using customized performance and resiliency policies. To avoid I/O bottlenecks and improve data velocity, we ensure careful task and data placement and explore customized I/O middleware policies.

Optimizing emerging cloud execution models. We propose retaining the attractive properties of FaaS execution but avoiding its overheads. We explore customized workflow performance and resiliency configurations that avoid this overhead.

References

Akers, S., et al. 2021. "Rapid and Flexible Segmentation of Electron Microscopy Data Using Few-Shot Machine Learning," *NPJ Computational Materials*, **7**(1), 187.

Friese, R. D., et al. 2020. "Effectively Using Remote I/O for Work Composition in Distributed Workflows," *Proceedings of the 2020 Institute of Electrical and Electronics Engineers International Conference on Big Data*.

Kilic, O. O., et al. 2022. "MemGaze: Rapid and Effective Load-Level Memory and Data Analysis," *Proceedings of the 2022 Institute of Electrical and Electronics Engineers International Conference on Cluster Computing*.

Olszta, M., et al. 2022. "An Automated Scanning Transmission Electron Microscope Guided by Sparse Data Analytics," *Microscopy and Microanalysis*, **28**(5), 1611–21.

Suetterlein, J., et al. 2019. "TAZeR: Hiding the Cost of Remote I/O in Distributed Scientific Workflows," *Proceedings of the 2019 Institute of Electrical and Electronics Engineers International Conference on Big Data*.

Accelerated Trait-Based Modeling of Biogenic Methane Dynamics Using Physics-Guided Machine Learning

Jinyun Tang, Qing Zhu, William J. Riley, Eoin Brodie

Lawrence Berkeley National Laboratory

Focal Areas

- Key uncertainties and knowledge gaps where new methodology, infrastructure, or technology can advance predictive understanding of the methane cycle.
- A solution to a key challenge in implementing AI approaches (e.g., improving uncertainty quantification, federated learning) across the biological and environmental science domains as it pertains to the methane cycle.

Science or Technological Challenge

Global CH_4 emissions are dominated by biogenic sources resulting from the interplay between production by methanogens and consumption by methanotrophs. While process-based models exist and have been applied for a long time, they frequently fail to accurately capture the response of net CH_4 emissions to variations in environmental factors such as temperature, moisture, and pH. The explicit representation of microbial dynamics has been suggested to improve these models. However, determining how much complexity should be represented in these microbial models is difficult because both CH_4 production and oxidation are carried out by diverse groups of microbes that interact and compete with each other. Trait-based modeling approaches have been proposed to represent the diversity of microbes within a microbial community and their effects on CH_4 biogeochemistry. However, this approach becomes challenging due to the large computational costs for parameterization when more microbes are represented. Moreover, the high computational costs make it challenging to incorporate empirical observations, constrain model parameterization, and quantify modeling uncertainty conditioned on current knowledge in measurements and modeling.

Rationale

The high computational costs associated with trait-based models are largely due to the high computing cost of the numerical solvers used to integrate the differential equations over space and time. This high computational cost, in turn, makes the process of improving model parameterization through model-data fusion more challenging, as it often requires numerous iterations of calibration with large ensemble simulations. Machine learning (ML) has demonstrated the potential to significantly speed up forward model simulations in areas such as weather and climate modeling and computational fluid dynamics (e.g., Scher and Messor 2019; Weyn et al. 2019; Kochkov et al. 2021). By creating high-fidelity surrogates of trait-based models using ML, we can accelerate both forward and calibration simulations, allowing for efficient quantification and reduction of parametric uncertainties. Furthermore, the ease of computing derivatives with respect to parameters makes it easier to fine-tune the ML-based surrogate models by incorporating a wider range of data. Finally, by building surrogates of trait-based models with different levels of complexity, we can quantify the relationship between model complexity and predictive uncertainty, and determine the optimal level of model complexity needed to predict future biogenic CH_4 dynamics.

Narrative

Our objective is to create a framework that combines (1) a synthetic database of CH_4-related biochemical variables generated by the microbial modules of EcoSIM [the land model being developed for BioEPIC, originally based on *ecosys* (Grant et al. 2017)] with varying levels of parameterization complexity in microbial dynamics; (2) machine learning surrogates trained using simulations from each complexity configuration; and (3) a model-data fusion framework that incorporates various observations to refine model parameters through the surrogates. To maintain interpretability, we will use a physics-guided machine learning approach, as demonstrated in our recent studies (Liu et al. 2022; Yuan et al. 2022). By repeatedly integrating these three components, we can continuously improve the microbial module of EcoSIM and its surrogates and assess the impact of observations on model predictions. Finally, the resulting observationally constrained surrogates will be used for ensemble extrapolation in various scenarios, quantifying uncertainty across different levels of complexity and determining the optimal complexity for robust CH_4 dynamics predictions.

References

Grant, R. F., et al. 2017. "Mathematical Modelling of Arctic Polygonal Tundra with *Ecosys*: 1. Microtopography Determines How Active Layer Depths Respond to Changes in Temperature and Precipitation," *JGR Biosciences* **122**(12), 3161-73. DOI:10.1002/2017JG004037.

Kochkov, D., et al. 2021. "Machine Learning-Accelerated Computational Fluid Dynamics," *Proceedings of the National Academy of Sciences of the United States of America* **118**(21), e2101784118. DOI:10.1073/pnas.2101784118.

Liu, L. C., et al. 2022. "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* **15**(7), 2839–58. DOI:10.5194/gmd-15-2839-2022.

Scher, S., G. Messori. 2019. "Weather and Climate Forecasting with Neural Networks: Using General Circulation Models (GCMs) with Different Complexity as a Study Ground," *Geoscientific Model Development* **12**(7), 2797–2809. DOI:10.5194/gmd-12-2797-2019.

Weyn, J. A., et al. 2021. "Sub-Seasonal Forecasting with a Large Ensemble of Deep-Learning Weather Prediction Models," *Journal of Advances in Modeling Earth Systems* **13**(7), e2021MS002502. DOI:10.1029/2021MS002502.

Yuan, K., et al. 2022. "Causality Guided Machine Learning Model on Wetland CH_4 Emissions Across Global Wetlands," *Agricultural and Forest Meteorology* **324**, 109115. DOI:10.1016/j.agrformet.2022.109115.

Integrating Genomic and Flux Data to Develop Predictive Models for Managing Methane Emissions

Clifton P. Bueno de Mesquita,[1] Susannah G. Tringe[1,2]

[1] DOE Joint Genome Institute, [2] Lawrence Berkeley National Laboratory

Focal Area

The focal area of this white paper is the importance of integrating high-potential datasets including soil genomic data, eddy covariance flux data, and remotely sensed flux data across spatial and temporal scales.

Science or Technological Challenge

It remains a challenge to develop a mechanistic and predictive model of methane fluxes across space and time that accurately predicts how fluxes respond to environmental changes and that could be used to develop and assess emission management strategies. Such a model must incorporate microbial metabolic and ecological processes occurring at local scales that ultimately scale up to landscape-level methane fluxes.

Rationale

Understanding microbial community taxonomic and functional composition has greatly increased our understanding of landscape-scale methane emission patterns across environmental gradients, yet predicting fluxes and how they respond to environmental change remains a major challenge (He et al. 2015; Hartman et al. 2017). Across the salinity gradient in the San Francisco Estuary, we used metagenomic sequencing to elucidate the involvement of multiple microbial functional guilds and decomposition processes that drove methane emissions that were highest in oligohaline wetlands but otherwise declined with increasing salinity (Hartman et al. 2023). A combination of metagenomic and metabolomic data also revealed halophilic methanogens contributing to the increased methane emissions observed in unrestored hypersaline salterns, a potential role for methane production by methylphosphonate-scavenging bacteria, and altered microbial community composition associated with lowered emissions after hydrologic restoration (Hartman et al. 2023). However, in a recent synthesis of methane flux and microbial data from coastal wetlands from four different sites across a wide geographic range, there were few consistencies in methane/salinity relationships and the variables driving them. Similar paired flux and microbial data from a greater number of sites would enable us to directly assess which environmental and microbial variables drive discrepancies among observed fluxes and environmental characteristics. This, in turn, could help predict the impact of changes in ecosystem management, restoration, or other interventions.

Narrative

A greater degree of integration between genomic and other omic data with methane flux data is needed at expanded spatial and temporal scales. A great deal of relevant data exist or are being generated, including land- and satellite-based methane monitoring data (e.g., Ameriflux and MethaneSAT), and metagenomic and metatranscriptomic data from soils and sediments [e.g., Integrated Microbial Genomes and Microbiomes database (Chen et al. 2021) and National Microbiome Data Collaborative], but have not been exploited to identify microbial-methane linkages. We propose leveraging these datasets, along with metadata repositories such as the Genomes OnLine Database (Mukherjee et al. 2021) and relevant ontologies to identify the environments, organisms, and metabolic pathways driving global methane emissions. One challenge in synthesizing these data is a lack of consistent metadata and paired microbial/methane measurements. We propose more soil sampling and microbial DNA sequencing efforts to be paired with already established methane monitoring efforts such as eddy covariance flux towers, especially in areas where such data are lacking, and the data integrated into appropriate repositories. Once enough data are generated from many sites, an analysis that synthesizes the data across sites and tests hypotheses about environmental/methane/microbial relationships would also benefit from machine learning techniques. Such techniques may include supervised machine learning models such as random forests, gradient boosting, support vector machines, ridge regression, and neural networks, or unsupervised methods (Ghannam and Techtmann 2021). These techniques will help reveal patterns in highly complex datasets comprising thousands of microbial taxa. Ultimately, these data could help develop a model of the methane cycle that explicitly includes microbial processes, similar to what has been done previously with soil carbon, arid ecosystems, and other climate models (Collins et al. 2008; Singh et al. 2010; Todd-Brown et al. 2012; Wieder et al. 2013). The model could then be used to help understand which interventions, out of a variety of different options (Valach et al. 2021), would lead to the greatest reductions in methane emissions.

References

Chen, I-M. A., 2021. “The IMG/M Data Management and Analysis System v.6.0: New Tools and Advanced Capabilities,” *Nucleic Acids Research* **49**: D751–D763.

Collins, S. L., et al. 2008. “Pulse Dynamics and Microbial Processes in Aridland Ecosystems,” *Journal of Ecology* **96**, 413–20.

Ghannam, R. B., and S. M. Techtmann. 2021. “Machine Learning Applications in Microbial Ecology, Human Microbiome Studies, and Environmental Monitoring,” *Computational and Structural Biotechnology Journal* **19**, 1092–1107.

Hartman, W. H., et al. 2017. “A Genomic Perspective on Stoichiometric Regulation of Soil Carbon Cycling,” *ISME Journal* **11**, 2652–665.

Hartman, W. H., et al. 2023. “Multiple Microbial Guilds Mediate Soil Methane Cycling Along a Wetland Salinity Gradient,” *ISME Journal.* In review.

He, S., et al. 2015. “Patterns in Wetland Microbial Community Composition and Functional Gene Repertoire Associated with Methane Emissions,” *mBio* **6**, e00066-15.

Mukherjee, S., et al. 2021. “Genomes OnLine Database (GOLD) v.8: Overview and Updates,” *Nucleic Acids Research* **49**: D723–D733.

Singh, B. K., et al. 2010. “Microorganisms and Climate Change: Terrestrial Feedbacks and Mitigation Options,” *Nature Reviews Microbiology* **8**: 779–90.

Todd-Brown, K. E. O., et al. 2012. “A Framework for Representing Microbial Decomposition in Coupled Climate Models,” *Biogeochemistry* **109**, 19–33.

Valach, A. C., et al. 2021. “Part III: Carbon Flux Trajectories and Site Conditions from Restored Impounded Marshes in the Sacramento-San Joaquin Delta,” In *Wetland Carbon and Environmental Management, Geophysical Monograph 267*, First Edition. Geophysical Monographs, American Geophysical Union/Wiley Publishers, Washington, D.C. www.wiley.com/en-us/Wetland+Carbon+and+Environmental+Management-p-9781119639282.

Wieder, W. R., et al. 2013. “Global Soil Carbon Projections are Improved by Modelling Microbial Processes. *Nature Climate Change* **3**, 909–12.

Zhou, J., et al. 2021. “Microbial Drivers of Methane Emissions from Unrestored Industrial Salt Ponds,” *ISME Journal* **16**: 284–95.

Uncertainty in Global Time-Resolved Methane Emissions from Aquatic Waterbodies

Jemma Stachelek,[1] Peter Regier,[2] Jon Schwenk,[1] Nicholas Ward[2]

[1]Los Alamos National Laboratory, [2]Pacific Northwest National Laboratory

Focal Areas

- Uncovering key uncertainties and knowledge gaps for predictive understanding of the methane cycle.
- Automated/real-time data capture to improve uncertainty quantification.

Science or Technological Challenge

Quantifying global methane emissions from lakes and reservoirs (hereafter referred to as waterbodies) is challenged both by (1) uncertainty in areal flux rates and (2) uncertainty in the distribution and magnitude of total waterbody area. This white paper is concerned with constraining estimates of the latter and propagating its associated uncertainties to global methane emissions calculations. To do this, we propose creating a data pipeline for retrieving time-resolved measurements of waterbody area and using these to update a methane uncertainty model, the architecture of which we describe below.

Rationale

Because we do not have a complete real-time census of all waterbodies, upscaling estimates of methane emissions from small waterbodies to broad spatial extents requires the use of waterbody size-abundance distributions rather than empirical measurements of area. Such waterbody-size abundance distributions are typically generated on an ad-hoc (i.e., one-off) basis that yields an over-exact estimate of total waterbody area reported with no uncertainty bounds (Keller et al. 2021). As an alternative to the typical approach, we propose a data assimilation workflow that combines Stachelek's (2023) Bayesian sensitivity analysis method with a dynamic data pipeline for retrieving waterbody areas from remote sensing imagery (Cooley et al. 2019). Our approach is capable of producing global waterbody nowcasts of methane emissions that include the uncertainty arising from dynamic area fluctuations. Not only does our approach avoid the necessarily static estimates derived from static waterbody databases, it avoids the need to continuously create massive global waterbody datasets (Pi et al. 2022) out of whole cloth. Instead, an initial estimate of methane emissions uncertainty is derived, which is then dynamically updated via a data pipeline that retrieves waterbody areas from remote sensing imagery.

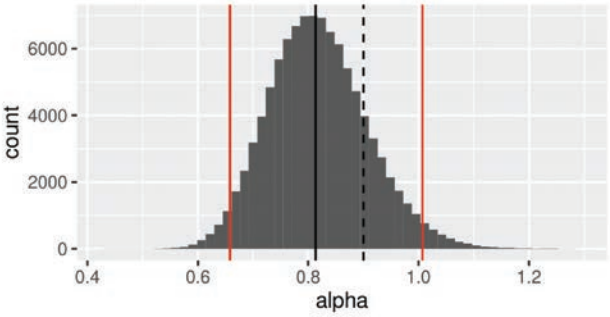


Fig. 1. Median (black line) and central 95% interval estimates of the Pareto shape parameter alpha (red lines) from a simulation. Here, the true alpha is 0.9. [Reprinted under a Creative Commons Attribution License (CC BY) from Stachelek, J. 2023. "Quantifying Uncertainty in Pareto Estimates of Global Lake Area," *Limnology and Oceanography: Methods* 21(3), 164–68. DOI:10.1002/lom3.10536.]

Narrative

The approach we describe below will help better define uncertainties in our predictive understanding of the methane cycle using advanced statistical tools and automated (near) real-time data capture. It involves two components:

Bayesian Sensitivity Analysis

Waterbody areas are typically treated as arising from a scale-invariant fractal generating process. This means that the number of waterbodies in one size class is proportional to the number of waterbodies in the preceding size class irrespective of their magnitudes (Stachelek 2023). The numerical form describing such a process is a power-law function. One of the statistical tools often used to model data that follow a power-law function is the Pareto distribution. The fit of any particular dataset to a Pareto distribution has associated uncertainty (see Fig. 1), which can be carried through to uncertainty in waterbody areas (Stachelek 2023), and we propose to carry this uncertainty forward even further to methane emissions calculations themselves (i.e., values across the posterior interval of the underlying parameters are used for calculation instead of a single posterior median).

Automated Data Capture

Even after accounting for uncertainty in total waterbody area on a static basis (Stachelek 2023), there remains a high degree of uncertainty with respect to dynamic fluctuations in waterbody area (Cooley et al. 2019; Pi et al. 2022). For example, new waterbodies are formed as a result of flooding, and old waterbodies disappear as a result of climate change and dam removal. Therefore, we propose a data pipeline which will retrieve raw remote sensing imagery and subject this imagery to water detection analysis, vectorization, and filtering for recurrency to exclude ephemeral and nonlake nonreservoir waterbodies. Limiting the pipeline to recurrent waterbodies will allow for temporal updating of an initial static uncertainty model (described above) and has the advantage of not requiring fully global processing. Rather, the model can be updated from limited portions of the globe as they become available in the data pipeline. When the automated data capture pipeline is fully integrated with the Bayesian sensitivity model, it will provide estimates of global methane emissions from inland waterbodies along with an associated uncertainty.

References

Cooley, S. W., et al. 2019. "Arctic-Boreal Lake Dynamics Revealed Using CubeSat Imagery," *Geophysical Research Letters* **46**, 2111–120. DOI:10.1029/2018GL081584.

Isikdogan, L. F., et al. 2019. "Seeing Through the Clouds with Deepwatermap," *IEEE Geoscience and Remote Sensing Letters* **17**(10), 1662–666. DOI:10.1109/LGRS.2019.2953261.

Keller, P. S., et al. 2021. "Global Carbon Budget of Reservoirs is Overturned by the Quantification of Drawdown Areas," *Nature Geoscience* **14**, 402–08. DOI:10.1038/s41561-021-00734-z.

Pi, X., et al. 2022. "Mapping Global Lake Dynamics Reveals the Emerging Roles of Small Lakes," *Nature Communications* **13**, 5777. DOI:10.1038/s41467-022-33239-3.

Planet Team. "Planet Application Program Interface: In Space for Life on Earth," https://api.planet.com/.

Stachelek, J., et al. 2022. "Identifying False Positive Lakes in Surface Water Detection," *HydroML Symposium 2022.*

Stachelek, J., 2023. "Quantifying Uncertainty in Pareto Estimates of Global Lake Area," *Limnology and Oceanography: Methods* **21**(3), 164–68. DOI:10.1002/LOM3.10536.

Estimation of Global Methane Soil Sink Using Multi-Source Datasets and Knowledge-Guided Machine Learning

Youmi Oh,[1] Licheng Liu,[2] Qing Zhu,[3] Gavin McNicol,[4] Zhenong Jin,[2] Sparkle Malone[5]

[1]NOAA Global Monitoring Laboratory, [2]University of Minnesota, [3]Lawrence Berkeley National Laboratory, [4]University of Illinois–Chicago, [5]Yale University

Focal Areas

- 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

There is a large spatial and temporal uncertainty in estimating global CH_4 soil oxidation, and reducing the uncertainty is important to reduce the bias in global CH_4 budgets. There is no study that uses advanced knowledge-guided machine learning (KGML) models to estimate global CH_4 soil sinks.

Rationale

CH_4 oxidation by microbes is the second largest sink in the global CH_4 budget, but its importance has been widely underestimated (Conrad 2009). Recent studies identified an overlooked CH_4 soil sink in diverse terrestrial ecosystems, such as Arctic tundra and forest, grassland, tropical savanna, and the Antarctic (Kato et al. 2011; Lau et al. 2015; Covey and Megonigal 2019; Ortiz et al. 2021). This soil sink has been attributed to high affinity methanotrophs (HAM), which grow on atmospheric CH_4 in dry mineral soils (Oh et al. 2020) and are highly temperature sensitive (Lau et al. 2015). We previously incorporated HAM into pan-Arctic CH_4 models and found that the soil sink could be twice the current estimate and will increase in the future due to a strong temperature sensitivity (Oh et al. 2020).

The underestimated global CH_4 soil sink can partly explain the discrepancy between the global CH_4 budget estimated by bottom-up mechanistic models and top-down atmospheric inversions (Jackson et al. 2020; Saunois et al. 2020). The mechanistic models overestimate the CH_4 budget by 175 $TgCH_4yr^{-1}$ when compared with atmospheric inversions, mostly due to high emission estimates from natural sources. The current estimation of global CH_4 soil sink is ~30 $TgCH_4yr^{-1}$ but with a huge uncertainty (7 to >100 $TgCH_4yr^{-1}$) from previous studies (Dutaur and Verchot 2007; Smith et al. 2000).

Long-term trends in the global CH_4 soil sink are also highly uncertain. A meta-analysis study showed that CH_4 oxidation from global forest soils decreased by 77% from 1988 to 2015 and this change was driven by an increase in precipitation (Ni and Groffman 2018). However, this argument has been challenged due to biased CH_4 oxidation samples across wet/dry sites and years. In contrast, another meta-analysis study showed that CH_4 oxidation increases when precipitation increases for tropical, savanna, and boreal ecosystems (Gatica et al. 2020), and mechanistic models of global CH_4 soil sink show a long-term increasing trend due to increases in temperature and atmospheric CH_4 (Zhuang et al. 2013; Murguia-Flores et al. 2021). Accurately quantifying the size and trends in the soil sink is extremely important to reduce the bias in current and future global CH_4 budgets.

Multisource datasets are available at various scales and measured with different techniques (e.g., chamber, eddy covariance data, etc.). However, there is a lack of an effective way to extrapolate/upscale the site-level observations and knowledge to the global scale. Mechanistic modeling and machine learning (ML) approaches have been widely used to scale and quantify CH_4 fluxes to regional and global scales (Zhuang et al. 2013; Peltola et al. 2019; Kim et al. 2020; Murguia-Flores et al. 2021), but both approaches show their own limitations. Mechanistic modeling incorporates scientific knowledge into the upscaling, yet large uncertainties arise if location- and vegetation-specific parameters are not calibrated properly, or if the underlying mechanisms are oversimplified or incompletely represented (Oh et al. 2020). The data-driven ML is increasingly popular in Earth system sciences due to its potential for high computational efficiency and accuracy (Rasp et al. 2018; Peltola et al. 2019; Jung et al. 2020; Kim et al. 2020; Irrgang et al. 2021).

However, existing ML models suffer from out-of-sample prediction failure in the absence of large training datasets and the results of ML models are often uninterpretable due to the black box use (Hutson 2022).

The growing field of KGML (Karpatne et al. 2022; Willard et al. 2023) provides a promising hybrid modeling method that combines the advantages of mechanistic models, ML models, and multisource datasets. KGML has successfully modeled certain Earth systems in which dynamic processes are well represented by established governing equations, such as in hydrology and atmospheric sciences (Read et al. 2019; Beucler et al. 2021; Irrgang et al. 2021; ElGhawi et al. 2023; Kraft et al. 2022; Willard et al. 2023). However, biogeochemical processes would be mathematically highly nonlinear and complicated. Unlike atmospheric systems, soil processes in terrestrial ecosystems cannot be directly observed by remote sensing, and *in situ* measurements are often expensive and limited. In this white paper, we propose to develop a novel KGML approach to incorporate biogeochemical knowledge into ML and effectively assimilate multisource measurements to capture complex soil CH_4 oxidation processes.

Narrative

We propose to develop a KGML framework that incorporates known biogeochemical principles into ML to improve model training, interpretability, and accuracy across global spatial and monthly-to-interannual temporal variability. Mechanistic models will be used as scientific foundations to develop the KGML hierarchical structure (Khandelwal et al. 2020; Liu et al. 2021, 2022) and generate millions of synthetic data for KGML pretraining (Read et al. 2019; Kraft et al. 2022). We will build separate ML modules for soil thermal, hydrological, and biogeochemical processes and an overarching model structure to link the submodules (Liu et al. 2021, 2022). To capture the complex biogeochemical processes, we will investigate advanced ML methods such as recurrent, convolutional, and graph neural networks (RNN, CNN, and GNN, respectively), as well as more recent techniques such as attention models and transformers (Vaswani et al. 2017; Dosovitskiy et al. 2020). The key biogeochemical constraints (e.g., CH_4 substrate and soil temperature influences) will be carefully embedded into the cost function using known principles or empirical functions (e.g., Michaelis-Menten kinetics and Q_{10} equation) as extra knowledge-guided losses (Read et al. 2019; Khandelwal et al. 2020; Liu et al. 2021; Yuan et al. 2022). The developed KGML will be further trained/validated with multi-source observations of soil CH_4 sink. Satellite remote sensing data or reanalysis data will be assimilated to constrain the KGML model internal processes (e.g., soil hydrology process intermediate output) to better capture the temporal and spatial heterogeneity.

References

Beucler, T., et al. 2021. "Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems," *Physical Review Letters* **126**(9), 098302.

Conrad, R. 2009. "The Global Methane Cycle: Recent Advances in Understanding the Microbial Processes Involved," *Environmental Microbiology Reports* **1**, 285–92.

Covey, K. R., and J. P. Megonigal. 2019. "Methane Production and Emissions in Trees and Forests," *New Phytologist* **222**, 35–51.

Dosovitskiy, A., et al. 2020. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv* preprint. DOI:10.48550/arXiv.2010.11929.

Dutaur, L., and L. V. Verchot 2007. "A Global Inventory of the Soil CH_4 Sink," *Global Biogeochemical Cycles* **21**(4). DOI:10.1029/2006GB002734.

ElGhawi, R., et al. 2023. "Hybrid Modeling of Evapotranspiration: Inferring Stomatal and Aerodynamic Resistances Using Combined Physics-Based and Machine Learning," *Environmental Research Letters* **18**, 034039. DOI:10.1088/1748-9326/acbbe0.

Gatica, G., et al. 2020. "Environmental and Anthropogenic Drivers of Soil Methane Fluxes in Forests: Global Patterns and Among-Biomes Differences," *Global Change Biology* **26**(11), 6604–15. DOI:10.1111/gcb.15331.

Hutson, M. 2022. "Taught to the Test," *Science* **376**(6593) 570–73. DOI:10.1126/science.abq7833.

Irrgang, C., et al. 2021. "Towards Neural Earth System Modelling by Integrating Artificial Intelligence in Earth System Science," *Nature Machine Intelligence* **3**, 667–74.

Jackson, R. B., et al. 2020."Increasing Anthropogenic Methane Emissions Arise Equally from Agricultural and Fossil Fuel Sources," *Environmental Research Letters* **15**(7), 071002.

Jung, M., et al. 2020. "Scaling Carbon Fluxes from Eddy Covariance Sites to Globe: Synthesis and Evaluation of the FLUXCOM Approach," *Biogeosciences* **17**(5) 1343–365.

Karpatne, A., et al., Eds. 2022. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data.* CRC Press, Boca Raton, FL.

Kato, T., et al. 2011. "Spatial Variability of CH_4 and N_2O Fluxes in Alpine Ecosystems on the Qinghai–Tibetan Plateau," *Atmospheric Environment* **45,** 5632–639. DOI:10.1016/j.atmosenv.2011.03.010.

Khandelwal, A., et al. 2020. "Physics Guided Machine Learning Methods for Hydrology," *arXiv* preprint. DOI:10.48550/arXiv.2012.02854.

Kim, Y., et al. 2020. "Gap-Filling Approaches for Eddy Covariance Methane Fluxes: A Comparison of Three Machine Learning Algorithms and a Traditional Method with Principal Component Analysis," *Global Change Biology* **26**(3), 1499–1518.

Kraft, B., et al. 2022. "Towards Hybrid Modeling of the Global Hydrological Cycle," *Hydrology and Earth System Sciences* **26**(6), 1579–1614 . DOI:10.5194/hess-26-1579-2022.

Lau, M. C. Y., et al. 2015. "An Active Atmospheric Methane Sink in High Arctic Mineral Cryosols," *ISME Journal* **9**, 1880–891.

Liu, L., et al. 2021. "Estimating the Autotrophic and Heterotrophic Respiration in the U.S. Crop Fields using Knowledge Guided Machine Learning," *ESSOAr*. DOI:1002/esso10ar.10509206.1.

Liu, L., et al. 2022. "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* **15**(7), 2839–58 DOI:10.5194/gmd-15-2839-2022.

Murguia-Flores, F., et al. 2021. "Uptake of Atmospheric Methane by Soil From 1900 to 2100," *Global Biogeochemical Cycles* **35**(7), e2020GB006774.

Ni, X., and P. M. Groffman 2018. "Declines in Methane Uptake in Forest Soils," *PNAS* **115**(34), 8587–90. DOI:10.1073/pnas.1807377115.

Oh, Y., et al. 2020. "Reduced Net Methane Emissions Due to Microbial Methane Oxidation in a Warmer Arctic," *Nature Climate Change* **10**, 317–21. DOI:10.1038/s41558-020-0734-z.

Ortiz, M., et al. 2021. "Multiple Energy Sources and Metabolic Strategies Sustain Microbial Diversity in Antarctic Desert Soils," *Proceedings of the National Academy of Sciences* **118**(45), e2025322118.

Peltola, O., et al. 2019. "Monthly Gridded Data Product of Northern Wetland Methane Emissions Based on Upscaling Eddy Covariance Observations," *Earth System Science Data* **11**(3), 1263–89.

Rasp, S., et al. 2018. "Deep Learning to Represent Subgrid Processes in Climate Models," *PNAS* **115**(39), 9684–89.

Read, J. S., et al. 2019. "Process-Guided Deep Learning Predictions of Lake Water Temperature," *Water Resources Research* **55**(11), 9173–90. DOI:10.1029/2019wr024922.

Riley, W. J., et al. 2011. "Barriers to Predicting Changes in Global Terrestrial Methane Fluxes: Analyses Using CLM4Me, a Methane Biogeochemistry Model Integrated in CESM," *Biogeosciences* **8**(7), 1925–953. DOI:10.5194/bg-8-1925-2011.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12(3),** 1561–1623.

Smith, K. A., et al. 2000."Oxidation of Atmospheric Methane in Northern European Soils, Comparison with other Ecosystems, and Uncertainties in the Global Terrestrial Sink," *Global Change Biology* **6**(7), 791–803.

Vaswani, A., et al. 2017. "Attention is All you Need," *Advances in Neural Information Processing Systems* **30**, 6000–010.

Willard, J., et al. 2023. "Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems," *ACM Computing Surveys* **55**(4), 1–37. DOI:10.1145/3514228.

Yuan, K., et al. 2022. "Causality Guided Machine Learning Model on Wetland CH_4 Emissions Across Global Wetlands," *Agricultural Forest Meteorology* **324**, 109115. DOI:10.1016/j.agrformet.2022.109115.

Zhuang, Q., et al. 2013. "Response of Global Soil Consumption of Atmospheric Methane to Changes in Atmospheric Climate and Nitrogen Deposition," *Global Biogeochemical Cycles* **27**(3), 650–63.

Toward Spatiotemporally Resolved Methane Emissions for Modeling and Upscaling Research

Housen Chu[1] (hchu@lbl.gov), Gavin McNicol,[2] Qing Zhu,[1] Avni Malhotra,[3] Camilo Rey-Sanchez,[4] David Durden,[5] Stefan Metzger,[5,6]

[1]Lawrence Berkeley National Laboratory, [2]University of Illinois–Chicago, [3]Pacific Northwest National Laboratory, [4]North Carolina State University, [5]National Ecological Observatory Network, [6]University of Wisconsin–Madison

Focal Areas

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

Science or Technological Challenge

Networks of eddy covariance towers, such as AmeriFlux and FLUXNET, provide large datasets of ecosystem energy, water, and carbon fluxes, enabling upscaling from sparse observations to regional/global flux predictions (Jung et al. 2019). Recently, the FLUXNET-CH4 initiative harmonized methane flux data from 81 sites, primarily wetlands, aiming to provide bottom-up upscaled methane fluxes (Delwiche et al. 2021). While eddy covariance data are recognized for their rich temporal information, their spatial dynamics are often overlooked and remain a primary source of uncertainties (Xu, F., et al. 2017; Metzger 2018; Chu et al. 2021). Briefly, the source area contributing to the flux at each time (i.e., flux footprint) varies depending on the effective measurement height, underlying surface characteristics, and turbulent state of the atmosphere. This spatiotemporal dynamic nature poses a critical challenge, particularly at sites with heterogeneous underlying sources/sinks such as wetlands. Hot spots and moments of methane emissions can form due to fine-scale variability driven by subsurface biogeochemistry, hydrologic gradient, salinity, nutrient availability, soil characteristics, vegetation types, and microtopography. The spatiotemporally dynamic footprints and sources/sinks jointly could lead to ~14%–25% biases (Matthes et al. 2014; Rey-Sanchez et al. 2018; Tuovinen at al. 2019) in area-integrated methane emissions and up to 83% in an extreme case (Morin et al. 2017). While recognizing the spatiotemporal dynamics, it remains challenging to incorporate the footprint information into the modeling and upscaling framework.

Rationale

Numerous research studies have attempted to address this “footprint” challenge, mostly in single-site studies with specific considerations of site characteristics and underlying processes. Attempts also varied regarding additional data requirements [e.g., chamber flux (Rey-Sanchez et al. 2018), paired towers (Matthes et al. 2014), spatial surface characteristics (Xu, K., et al. 2017; Tuovinen at al. 2019), wavelet-based flux calculation (Xu, K., et al. 2017)] and core model types/structures [e.g., biophysical (Duman and Schäfer 2018), statistical model (Matthes et al. 2014; Xu, F., et al 2017; Tuovinen at al. 2019; Levy et al. 2020), vegetation index-based (Ran et al. 2016), machine learning (Xu, K., et al. 2017), hybrid approach (Xu, K., et al. 2017; Metzger 2018; Wiesner et al. 2022)]. While deemed promising individually, there have been limited attempts to benchmark the proposed approaches across sites, particularly for methane fluxes. We attributed the research latency to the following challenges. First, flux-decomposing research mostly began with pre-identified/hypothesized hot spots or spatial gradients. Yet, eddy covariance flux data contain rich temporal information reflecting a combination of complex and dynamic processes over different timescales. Thus, spatial flux information is masked and confounded by temporal variability, hindering spatially explicit investigations. Second, the additional data requirements remain a significant hurdle. For example, very few eddy covariance wetland sites have co-located, continuous, and representative chamber measurements (e.g., over vegetation, soil, and open water; Määttä et al. In preparation) that help constrain or validate the flux decomposition. Also, fine-resolution (both temporal and spatial) surface characteristics, such as vegetation indices, surface temperature, and soil moisture, are rarely available. Third, most approaches require prior knowledge of the methane flux's controlling mechanisms, which might vary across wetlands or land cover types within the site, further complicating the generalization of approaches across sites. A few studies have proposed a machine learning–based approach to derive environmental response functions, which combine observations, processes, and data mining to express the spatiotemporal flux (Xu et al. 2017; Metzger 2018). This approach uses a universal model across a site's flux footprints and reconciles observed spatiotemporal dynamics based on temporal and spatial covariates. A hybrid approach built upon this framework was proposed to incorporate the machine-learned spatiotemporal dynamics into a process-based model (Wiesner et al. 2022). It extracts multidimensional processes from the environment constrained by knowledge-based processes and creates georeferenced maps and process benchmarks for geostatistics, model evaluation, and upscaling.

Narrative

We propose future synthesis to build a robust, scalable workflow to decompose methane fluxes measured using the eddy covariance technique, producing the spatiotemporally resolved, debiased ecosystem methane emissions for modeling and upscaling research. Machine learning can help fill the workflow's technical and data gaps discussed earlier. First, a recent study proposed a simplistic approach to derive a hot spot flux map based mainly on eddy covariance data (Rey-Sanchez et al. 2022). The method can better identify and delineate potential hot spots and their flux contributions when paired with a knowledge-based land cover map. Machine learning–based classification can be a surrogate or a means for accurate, fine-scale wetland land-cover classifications across sites (Palace et al. 2018). Second, several new constellations of satellites (e.g., PlanetScope and HydroSat) are becoming available and shedding light on fine spatiotemporal surface characteristics in the foreseeable future. Machine learning approaches can help generate robust, downscaled, fine-resolution surface characteristics before the desired retrievals become available (Greifeneder et al. 2021). We also advocated future efforts to collect and synthesize chamber fluxes for providing ground-truth validation (Määttä et al. In preparation). Third, while machine learning has demonstrated the potential to learn and simulate the spatiotemporal flux dynamics, many previous studies still adopted a process-based core model for decomposing the spatial fluxes. We suggested that machine learning methods can serve as a data-exploring tool to detect relationships and interactions that help unveil new microbiological and biogeochemical processes. Further research should also explore the potential of a hybrid modeling approach, taking advantage of process-based and machine learning models, attributing the spatial variability, and informing site design and validation studies.

References

Chu, H., et al. 2021. "Representativeness of Eddy-Covariance Flux Footprints for Areas Surrounding AmeriFlux Sites," *Agricultural and Forest Meteorology* 301–02, 108350.

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–689.

Duman, T., and K. V. R. Schäfer. 2018. "Partitioning Net Ecosystem Carbon Exchange of Native and Invasive Plant Communities by Vegetation Cover in an Urban Tidal Wetland in the New Jersey Meadowlands (USA)," *Ecological Engineering* 114, 16–24.

Greifeneder, F., et al. 2021. "A Machine Learning-Based Approach for Surface Soil Moisture Estimations with Google Earth Engine," *Remote Sensing* **13**, 2099.

Jung, M., et al. 2019. "The FLUXCOM Ensemble of Global Land-Atmosphere Energy Fluxes," *Scientific Data* **6**(74).

Levy, P., et al. 2020. "Inference of Spatial Heterogeneity in Surface Fluxes from Eddy Covariance Data: A Case Study from a Subarctic Mire Ecosystem," *Agricultural and Forest Meteorology* 280, 107783.

Määttä T., et al. "Spatial Heterogeneity Dictates Coherence Between Eddy Covariance- and Chamber-Based CH_4 Measurements Across Multiple Wetland Sites." In preparation.

Matthes, J. H., et al. 2014. "Parsing the Variability in CH_4 Flux at a Spatially Heterogeneous Wetland: Integrating Multiple Eddy Covariance Towers with High-Resolution Flux Footprint Analysis," *Journal of Geophysical Research: Biogeosciences* **119**, 2014JG002642.

Metzger, S. 2018. "Surface-Atmosphere Exchange in a Box: Making the Control Volume a Suitable Representation for *In Situ* Observations," *Agricultural and Forest Meteorology*, **255**, 68–80.

Morin, T. H., et al. 2017. "Combining Eddy-Covariance and Chamber Measurements to Determine the Methane Budget from a Small, Heterogeneous Urban Floodplain Wetland Park. *Agricultural and Forest Meteorology* **237–238**, 160–70.

Palace, M., et al. 2018. "Determining Subarctic Peatland Vegetation Using an Unmanned Aerial System (UAS)," *Remote Sensing* **10**(9), 1498.

Ran, Y., et al. 2016. "Spatial Representativeness and Uncertainty of Eddy Covariance Carbon Flux Measurements for Upscaling Net Ecosystem Productivity to the Grid Scale," *Agricultural and Forest Meteorology* **230**, 114–27.

Rey-Sanchez, A. C., et al. 2018. "Determining Total Emissions and Environmental Drivers of Methane Flux in a Lake Erie Estuarine Marsh," *Ecological Engineering* **114**, 7–15.

Rey-Sanchez, C., et al. 2022. "Detecting Hot Spots of Methane Flux Using Footprint-Weighted Flux Maps," *Journal of Geophysical Research: Biogeosciences* **127**(8), e2022JG006977.

Tuovinen, J. P., et al. 2019. "Interpreting Eddy Covariance Data from Heterogeneous Siberian Tundra: Land-Cover-Specific Methane Fluxes and Spatial Representativeness," *Biogeosciences* **16**(2), 255–74.

Wiesner, S., et al. 2022. "Quantifying the Natural Climate Solution Potential of Agricultural Systems by Combining Eddy Covariance and Remote Sensing," *Journal of Geophysical Research Biogeosciences* **127**(9).

Xu, F., et al. 2017. "Area-Averaged Evapotranspiration Over a Heterogeneous Land Surface: Aggregation of Multi-Point EC Flux Measurements with a High-Resolution Land-Cover Map and Footprint Analysis," *Hydrology and Earth System Sciences* **21**(8), 4037–051.

Xu, K., et al. 2017. "Upscaling Tower-Observed Turbulent Exchange at Fine Spatio-Temporal Resolution Using Environmental Response Functions," *Agricultural and Forest Meteorology* **232**, 10–22. DOI:10.1016/j.agrformet.2016.07.019.

Scaling Genes to Global Methane Modeling Through Artificial Intelligence

Xiaofeng Xu[1] (xxu@sdsu.edu), Jorge Rodrigues[2] (jmrodrigues@ucdavis.edu)

[1]San Diego State University, [2]University of California Davis

Focal Areas

- 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 combining data across spatial or temporal scales or scientific domains may lead to new scientific insights, either within or across fields.

Science or Technological Challenge

Methane (CH_4) is produced at the microscale, while policymaking relies on macro-scale CH_4 information. The ability to understand and predict CH_4 cycling across distinct scales is essential but remains a grand challenge. Microbiologists have produced a huge amount of metagenomic data on CH_4-relevant functional genes (Freitag and Prosser 2009; Kroeger et al. 2020), which can inform the individual CH_4 processes at the microscale. Correspondingly, land surface CH_4 flux data can infer the CH_4 cycling at an ecosystem level, and atmospheric CH_4 concentration fluctuation implies CH_4 cycling at a regional scale. Building a multiscale modeling capability can benefit from integrating data obtained at various scales but is particularly limited by the modeling capability for assimilating metagenomic data on methanogenesis and methanotrophy (Xu et al. 2015; Sihi et al. 2021).

Rationale

Physical and chemical processes are predictable with high confidence, while biological processes remain challenging for accurate prediction which heavily relies on massive datasets and robust models. Methane modeling techniques have been developed and applied for more than 40 years, and more than 40 CH_4 models have been developed from 1987 to 2016 (Xu et al. 2016). Considering the CH_4 models developed in the past 6 years (Song et al. 2020; Ricciuto et al. 2021; Sihi et al. 2021; Yuan et al. 2021), substantial progress has been made in modeling CH_4 cycling processes. Meanwhile, progress has been made in gathering large datasets of functional genes encoding proteins responsible for CH_4 production and oxidation in various biomes. Knowledge gaps still exist in three aspects: (1) identifying the driving factors for microbial mechanisms associated with CH_4 production and oxidation, (2) connecting the microscale processes with large-scale CH_4 fluctuation with high predictability, and (3) parameterizing multiscale CH_4 models for simulating CH_4 cycling within an Earth system modeling framework. Because increased data cumulation did not bring a significant improvement in our confidence in predicting CH_4 fluxes in terrestrial ecosystems and in aquatic ecosystems as well, we propose that integrating microbial genomic data with ecosystem-level measurements through advanced artificial intelligence would significantly improve our predictability of CH_4 flux. This should be an achievable key task for the next 10 years. The mechanistic modeling approach carries the advantage of representing each CH_4 process individually while allowing for the integration of multiple sources of data (Xu et al. 2016). Advanced artificial intelligence (AI) algorithms can also be used for identifying genetic makers with direct association with CH_4 emissions within large metagenomic datasets (Khan et al. 2023) but not previously linked to the processes of methanogenesis and methanotrophy. These agnostic approaches being performed through AI are anticipated to better support the model parameterization and application in predicting CH_4 cycling.

Narrative

In order to develop robust predictability, the research community needs to enhance collaborative research for CH_4 modeling on and across three scales: (1) at the microscale, where different microbial processes are occurring to understand hot moments of emissions; (2) at the ecosystem scale, where CH_4 emissions are being measured to capture ecosystem drivers of methanogenesis and methanotrophic processes and validate models; and (3) at the regional/global scale, to upscale and predict changes over time. AI serves as a powerful tool to expedite the development of modeling capability by assisting in distilling information from massive data and further supporting model development and application. Specifically, we envision in-depth AI approaches to be used in three areas.

1. AI assistance in processing and integrating micro-scale meta-omics (metagenomics, metatranscriptomics, metaproteomics, and metabolomics) data with CH_4 models. Massive metagenomic data have been produced, but specific drivers of biological processes are challenging to retrieve. AI can be a powerful tool for understanding microbial physiology that is fundamental for methane production and oxidation processes occurring at the microscale. A study has applied a multifactorial strategy of deep sequencing and a machine learning approach to compare taxonomic differences and generated metabolic maps with differential representations of genes involved in the cycling of nutrients and CH_4 in forest and pasture soils in the Amazon Forest (Khan et al. 2023). This is an area that deserves further exploration as datasets have already been collected.

2. AI can assist in building ecosystem-level predictability based on plot-level observational data and micro-scale meta-omic datasets. Our group is working on a project to integrate metagenomic data with an ecosystem model to better parameterize the model for simulating individual CH_4 production processes rather than solely focusing on land surface CH_4 flux (Zuo et al. 2023). Our AI approach assists with model parameterization on meta-omic data and ecosystem-level CH_4 flux.

3. Enhance the Earth system model by including a microbial functional group-based CH_4 module with the capability of assimilating data of functional genes, ecosystem level CH_4 flux, and atmospheric CH_4 concentration. AI algorithms can be used to improve model efficiency and data assimilation.

References

Freitag, T. E., and J. I. Prosser. 2009. "Correlation of Methane Production and Functional Gene Transcriptional Activity in a Peat Soil," *Applied and Environmental Microbiology* **75**(21), 6679–687.

Khan, M. A. W., et al. 2023. "Amazonian Soil Metagenomes Indicate Different Physiological Strategies of Microbial Communities in Response to Land Use Change," *Applied Environmental Microbiology*. Under revision.

Kroeger, M. E., et al. 2020. "Rainforest-to-Pasture Conversion Stimulates Soil Methanogenesis Across the Brazilian Amazon," *The ISME Journal* **15**, 658–72.

Ricciuto, D., et al. 2021. "An Integrative Model for Soil Biogeochemistry and Methane Processes: I. Model Structure and Sensitivity Analysis," *Journal of Geophysical Research-Biogeosciences* **126**(8), e2019JG005468. DOI:10.1029/2019JG005468.

Sihi, D., et al. 2021. "Representing Methane Emissions from Wet Tropical Forest Soils Using Microbial Functional Groups Constrained by Soil Diffusivity," *Biogeosciences* **18**(5),1769–786.

Song, C., et al. 2020. "A Microbial Functional Group-Based CH_4 Model Integrated into a Terrestrial Ecosystem Model: Model Structure, Site-Level Evaluation, and Sensitivity Analysis," *Journal of Advances in Modeling Earth Systems* **12**(4), e2019MS001867.

Xu, X., et al. 2015. "A Microbial Functional Group Based Module for Simulating Methane Production and Consumption: Application to an Incubation Permafrost Soil," *Journal of Geophysical Research-Biogeosciences* **120**(7), 1315–333.

Xu, X., et al. 2016. "Review and Synthesis: Four Decades of Modeling Methane Cycling Within Terrestrial Ecosystems," *Biogeosciences* **13**(12), 3735–755.

Yuan, F., et al. 2021. "An Integrative Model for Soil Biogeochemistry and Methane Processes: II. Warming and elevated CO_2 Impacts on Peatland CH_4 Emission," *Journal of Geophysical Research-Biogeosciences* **126**(8), e2020JG005963. DOI:10.1029/2020JG005963.

Zuo, Y., et al. 2023. "Metagenomic Data for Improving Ecosystem Model in Simulating Methane Cycling in Temperate Wetland," *Journal of Advances in Modeling Earth Systems.* In preparation.

A Machine Learning Data Assimilation Method to Improve Lake Methane Prediction

Zeli Tan,[1] Johnny Li,[2] Hoang Viet Tran,[1] Dalei Hao[1]

[1]Pacific Northwest National Laboratory, [2]University of Idaho

Focal Area

The proposed research will demonstrate the potential to assimilate satellite and unmanned aerial vehicle (UAV) data into a lake methane model to improve the accuracy of lake methane prediction in the temperate climate regions of the Northern Hemisphere.

Science or Technological Challenge

Methane emissions from lakes are one of the largest natural methane sources, comprising as much as one-third of global methane emissions based on a recent estimate (Rosentreter et al. 2021). Lakes are prominent landscape elements in the temperate regions of the Northern Hemisphere. Due to climate warming, methane emissions from northern lakes are expected to rise rapidly in this century (Tan and Zhuang 2015). However, our capability to model methane emissions from northern lakes is still very limited. First, many related lake physical processes, such as water mixing and ice phenology, have not been well constrained in the current large-scale lake models (Guseva et al. 2020). Second, lake primary production not only provides labile organic carbon for methanogens to produce methane in anoxic conditions but also fuel methanogenesis in oxygen-rich conditions. It is found to be a critical factor for methane emissions from northern lakes but hasn't been well represented in the current large-scale lake models (West et al. 2016). Third, satellite-based methane data do not have sufficient signal-to-noise ratios and spatiotemporal resolutions to detect methane plumes from lake surface (Tan et al. 2016).

Rationale

Although data assimilation is widely used to improve the performance of numerical models, there are only limited applications for methane models. This is mainly because traditional data assimilation (DA) methods, such as ensemble Kalman filter (EnKF), have high computational and implementation costs. Also, traditional DA methods are not efficient to assimilate different types of data. To bridge the gaps, we propose to develop a long short-term memory (LSTM) neural network–based DA method that harnesses quality, multisource satellite data of ice cover, lake surface water temperature, Secchi depth, chlorophyll, and high-quality UAV data of methane fluxes to: (1) optimize the lake model's parameters and (2) improve estimations of lake methane emissions. Compared to traditional DA methods, the proposed machine learning–based method will be computationally efficient, code-change free, and bias-proofed against ill-selected likelihood functions (Tsai et al. 2021). Once validated, the method can be extended to temperate regions and even the high-latitude regions of the Northern Hemisphere to constrain lake methane emissions from these regions. The effort will strongly benefit to the accomplishment of the Global Methane Pledge.

Narrative

We will develop a LSTM neural network–based DA method to assimilate satellite and UAV data to improve the estimation of lake methane emissions in temperate regions of the Northern Hemisphere.

Description of the machine learning–based DA method. To overcome the limitations of traditional DA methods, we will develop a LSTM-based surrogate model to efficiently assimilate different types of lake data. Specifically, the neural density estimator will adjust the prior distribution of model parameters using satellite observations and UAV measurements to provide desirable sets of model parameters (see Fig. 1). We will construct the DA framework in three steps:

1. First, on the study lake, we will run a small number of lake model ensembles and then train a LSTM surrogate model using climate inputs, simulated lake dynamics, and model parameters. This step makes the surrogate LSTM model learn the ice phenology, mixing, primary production, and methane flux physics of the process-based lake methane model.

2. Next, we will use a neural density estimator to assess the uncertainty of the LSTM ensembles based on satellite

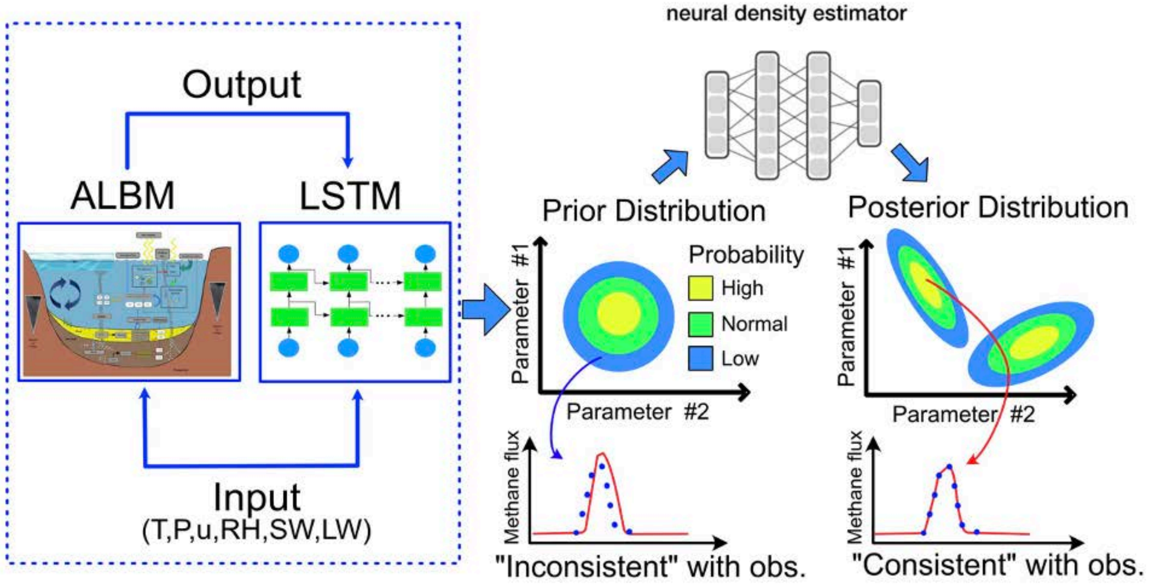


Fig. 1. Workflow of training a machine learning–based surrogate model of a 1-D lake methane model and using the surrogate model and a neural density estimator to assimilate satellite and unmanned aerial vehicle data to constrain simulated lake methane emissions. Climate inputs for Arctic Lake Biogeochemistry Model (ALBM) and long short-term memory (LSTM) are air temperature (T), precipitation (P), wind speed (u), relative humidity (RH), shortwave radiation (SW), and longwave radiation (LW).

and UAV observations and produce the posterior distribution of the parameters. The neural density estimator is a special approach to the inverse problem. Simulators use parameters (θ) to simulate data (Y). Inference goes the other direction, using observed data (Y_{True}) to get back the parameters (θ). Here, we will utilize the neural density estimator with satellite and UAV observations to infer better values for lake parameters with physical meaning.

3. Finally, lake parameters based on the posterior distributions will be fed to the LSTM surrogate model to produce estimations of lake thermal and methane dynamics that are close to observations. The estimated lake thermal and methane dynamics will be validated against *in situ* observations.

Description of the lake model. The Arctic Lake Biogeochemistry Model (ALBM) model is a 1-D process-based lake methane model developed by Dr. Zeli Tan (Tan et al. 2016). In this research, we will use the ALBM model to produce prior lake thermal and methane estimates for training the LSTM surrogate model. Here, we briefly describe the model processes and structure that are related to lake stratification simulations. ALBM is an integral energy lake model based on the Hostetler diffusivity parameterization, with depth-resolved 1-D water and sediment columns. Both the water and sediment columns have variable layer thickness, with thinner layers at surface to represent more intense thermal dynamics. In the model, lake methane emissions are governed by methane production, oxidation, and transport (via both diffusion and ebullition). The ALBM model has demonstrated good performance in simulating methane emissions of specific northern lakes (Tan et al. 2015; Guo et al. 2020).

Description of the satellite and UAV observations. We will use the satellite and UAV data of different lake thermal and biogeochemical variables together for data assimilation. All satellite data operations will be executed using Google Earth Engine. (1) For lake surface temperature, we will use the Advanced Very High Resolution Radiometer (AVHRR) sensor-based GloboLakes product for large lakes and the Landsat-8 thermal data for small lakes. (2) For ice cover, we will use the Advanced Microwave Scanning Radiometer (AMSR) sensor-based daily ice phenology data for large lakes and the Landsat-based ice phenology data for small lakes. (3) For Secchi depth, we will extract the values for the studied lakes from multiple satellite sensors, including Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat-8 Operational Land Imager (OLI), Sentinel-2 MultiSpectral Imager (MSI), and MEdium Resolution Imaging Spectrometer (MERIS), by adopting a well-established quasi-analytical algorithm. (4) For chlorophyll, we will use the data from the Landsat-8 OLI and the Ocean and Land Color Instrument (OLCI) aboard the Sentinel 3 satellite. (5) For methane fluxes, we will use drones carrying multimodal instruments to measure lake water surface temperature and map methane plumes close to sources and use coincident wind measurements to derive flux accurately and reliably. The approach will be developed at the University of Idaho.

References

Guo, M., et al. 2020. "Rising Methane Emissions from Boreal Lakes Due to Increasing Ice-Free Days," *Environmental Research Letters* **15**, 064008.

Guseva, S., et al. 2020. "Multimodel Simulation of Vertical Gas Transfer in a Temperate Lake," *Hydrology and Earth System Sciences* **24**, 697–715.

Rosentreter, J. A., et al. 2021. "Half of Global Methane Emissions Come from Highly Variable Aquatic Ecosystem Sources," *Nature Geoscience* **14**, 225–30.

Tan, Z., and Q. Zhuang. 2015. "Arctic Lakes are Continuous Methane Sources to the Atmosphere Under Warming Conditions," *Environmental Research Letters* **10**, 054016.

Tan, Z., et al. 2015. "Modeling Methane Emissions from Arctic Lakes: Model Development and Site-Level Study," *Journal of Advances in Modeling Earth Systems* **7**, 459–83.

Tan, Z., et al. 2016. "Inverse Modeling of Pan-Arctic Methane Emissions at High Spatial Resolution: What Can We Learn from Assimilating Satellite Retrievals and Using Different Process-Based Wetland and Lake Biogeochemical Models?" *Atmospheric Chemistry and Physics* **16**, 12649–666.

Tsai, W. P., et al. 2021. "From Calibration to Parameter Learning: Harnessing the Scaling Effects of Big Data in Geoscientific Modeling," *Nature Communications* **12**, 5988.

West, W. E., et al. 2016."Productivity and Morphometry Regulate Lake Contributions to Atmospheric Methane," *Limnology and Oceanography* **61**, S51–S61.

Physics-Guided Machine Learning of Wildfire Methane Emissions

Fa Li,[1] Min Chen,[1] Qing Zhu,[2] Kunxiaojia Yuan[2]

[1]University of Wisconsin–Madison, [2]Lawrence Berkeley National Laboratory

Focal Areas

- A solution to a key challenge in implementing advanced statistical approaches as it pertains to the methane cycle.
- Identifying high-potential datasets and how advanced statistical and numerical methods can be used to realize new scientific insights.
- How automated or real-time data capture and processing can be used to address issues of spatial and temporal heterogeneity and data sparsity.

Science or Technological Challenge

Global warming has significantly increased megafire risks, which are important sources of methane (such as the megafire in Indonesia in 1997; Turner et al. 2019), one of the most important greenhouse gasses. However, estimates of wildfire-induced methane emissions from either bottom-up (BU) or top-down (TD) models remain highly uncertain due to the imperfect model structures and limited data constraints. This would bring a big challenge to understanding and projecting future climate change.

Rationale

Methane emissions from wildfires are typically estimated using two primary approaches: BU and TD models. BU models rely on observations of combustion completeness (CC) and emission factors (EFs) for different plant functional types, which are often based on limited site observations. These estimates typically do not account for spatiotemporal variability in CC and EFs across different environments. However, BU models can provide high-resolution estimates of emissions.

On the other hand, TD models rely on atmospheric methane concentration measurements from towers or satellites, which are then used to estimate ground emissions from fires through inverse atmospheric transport modeling. TD models are typically more reliable at coarser spatial and temporal scales. TD models are often combined with BU models to estimate wildfire emissions, as BU models can provide important prior information for the TD approach.

To accurately estimate methane emissions from past and future fires, it is crucial to efficiently integrate data from various sources, such as ground, tower, and satellite observations, to constrain a coupled BU-TD model. Additionally, BU models need to be well parameterized for future projections. Traditionally, BU and TD models are parameterized separately, which requires a large number of time-consuming ensemble runs of the models. Furthermore, data assimilation algorithms, such as ensemble Kalman Filter, used to constrain these models often assume linear relationships between observations and model state variables (e.g., parameters), which may not be true.

In this white paper, we identified a few points where machine learning could be used to improve estimation and projection of methane emissions due to wildfires by integrating with BU and TD models.

Narrative

Machine learning (ML) helps solve the above-mentioned problems in several ways.

First, ML can be used to create surrogate models, which represent physics-guided ML that approximates the behavior of BU and TD models but with grand reduction of the computational cost, making data assimilation more efficient. For example, Zhu et al. (2022) built a deep neural network (DNN) scheme that surrogates the process-based wildfire model within the Energy Exascale Earth System Model (E3SM). The surrogate wildfire model successfully captured the observed regional burned area. Such models can be straightforwardly extended for the purpose of simulating methane emission due to fires.

Second, ML can also be used to learn and provide a more accurate representation of the relationship between observations and model variables. Satellite/tower observations usually only provide column methane concentrations at, near, or far from the locations where fires happen. The relationships between atmospheric column methane concentration, fire-emitted methane, EFs/CC are determined by the fire burning and atmospheric transport processes described in the BU and TD models, which are often highly nonlinear. ML can help identify and model these nonlinear relationships, which can improve the accuracy of data assimilation (Abarbanel et al. 2018). ML-based data assimilation is being used in weather forecasting but has not yet been applied for estimating methane emissions.

Additionally, ML can help with the selection and weighting of observations in the data assimilation process (Geer 2021). Traditional data assimilation methods often assume that all observations are equally important, but this may not always be the case. ML can help identify which observations are most important for improving model accuracy and assign appropriate weights to them.

References

Abarbanel, H. D. I., et al. 2018. “Machine Learning: Deepest Learning as Statistical Data Assimilation Problems,” *Neural Computation* **30**(8), 2025–55. DOI:10.1162/neco_a_01094.

Geer, A. J., 2021. “Learning Earth System Models from Observations: Machine Learning or Data Assimilation?” *Philosophical Transactions of the Royal Society A* **39**(2194). DOI:10.1098/rsta.2020.0089.

Turner, A. J., et al. 2019. “Interpreting Contemporary Trends in Atmospheric Methane,” *Proceedings of the National Academy of Sciences* **116**(8), 2805–13.

Zhu, Q., et al. 2022. “Building a Machine Learning Surrogate Model for Wildfire Activities Within a Global Earth System Model,” *Geoscientific Model Development* **15**(5), 1899–1911.

Characterizing Remotely Sensed CH_4 Through Biogenic and Anthropogenic Flux Source Attribution: An Ecosystem Embedding Approach

Dan J. Krofcheck (djkrofc@sandia.gov), Michael Nole

Sandia National Laboratories

Focal Areas

This work critically contributes to four core challenges associated with CH_4 monitoring and prediction:

- The development of a flexible modeling framework for upscaling biogenic CH_4 flux predictions.
- The improvement of existing inventory and emissions factor-based oil and gas (O&G) sector production models.
- Implementation of state-of-the-art AI for the uncertainty-aware and climate-responsive scaling of both anthropogenic and biogenic CH_4 sources.
- Incorporation into a top-down detection and source attribution remote sensing framework.

Science or Technological Challenges

This research program is multifaceted, with separate bottom-up and top-down research challenges. At its core is the comparison of models with observations, using instrumented research sites as a source of measurements. Both biogeochemical and O&G-centric models have data fusion and forecasting challenges, made more complex when paired with the requirement that the models are scaled spatially and interrogated using remote sensing CH_4 platforms.

Biogenic Focus

Significant effort has been made over the past decade to understand the mechanisms driving the generation and release of biogenic CH_4 from the land surface, especially in regions considered particularly sensitive to changes in climate like the Arctic. Site-specific validation studies have linked physical processes at the surface/subsurface level with local CH_4 flux measurements (e.g., chamber, flux tower) but remaining challenges include:

- Lack of closure between bottom-up and top-down CH_4 emissions measurements and insufficient measurements at the flux tower scale and in the shoulder seasons.
- Complexity of the processes involved not lending themselves to scaling due to computational cost or lack of ability to constrain models.
- Absence of a flexible framework to forward models across ecotypes or future ecotype transitions honoring uncertainty in future changes to Arctic permafrost.

Narrative and Rationale

Monitoring and predicting CH_4 emissions is an emerging challenge in the global biosphere-atmosphere flux community, the success of which will have significant impacts on our ability to constrain associated uncertainty and propose steps to mitigate runaway climatic change. CH_4 is produced through biogenic (natural) and anthropogenic (human-caused) sources and any attempts to characterize CH_4 remotely cannot inherently discriminate between the two sources. However, biogenic and anthropogenic mechanisms of CH_4 fluxes have very different abiotic drivers, with substantially variable responses to future climates and policy intervention efforts. For instance, biogenic CH_4 fluxes are ecotype-dependent and are controlled to varying degrees by surface temperature, moisture content, precipitation, leaf area index, lateral subsurface fluxes, organic matter composition, and soil physical properties, among other factors.

Anthropogenic sources of CH_4 flux are dominated by O&G infrastructure, with complex and poorly constrained understanding of how CH_4 emission from O&G varies as a function of atmospheric conditions and hardware state variables (e.g., component type, time since installation, time since maintenance, etc.). Current bottom-up modeling approaches subsume these mechanistic relationships using emissions factors and scaling spatially as a function of component composition, resulting in massive and poorly constrained uncertainties that are static with respect to climate. Our research here directly contributes to improved estimates and scaling of these anthropogenic fluxes through direct observations made using eddy covariance.

CH_4 flux measurements collected across networks of eddy covariance flux tower sites are a massively underleveraged source of direct ground-atmosphere fluxes of CH_4 that can be described as a function of biotic and anthropogenic state variables in response to changes in abiotic drivers. These CH_4 flux response functions will play a critical role in scaling local measurements to landscape and regional scales.

Our integrated CH_4 monitoring and decision framework combines bottom-up estimates of CH_4 emissions from biogenic and anthropogenic sources with top-down measurements from satellites and aerial platforms, using eddy covariance as a systems integration lens. Specifically, using sequence transformers used for language modeling to create an ecosystem-embedding model for the terrestrial fluxes of carbon, water, and energy in a general way, with specific inclusion of terrestrial sources of CH_4. This ecosystem embedding approach learns the relationship between abiotic drivers and CH_4 flux as a function of remotely retrievable state variables of the system. By describing these state variables specifically in terms of anthropogenic parameters (e.g., O&G infrastructure databases) and biogenic parameters (e.g., vegetation type, leaf area index), we can dramatically improve our ability to generate bottom-up emissions estimates, with direct biogenic or anthropogenic source attribution. Ultimately, this capability is designed to operate in concert with space-borne and aerial-gridded estimates of CH_4 concentration and will permit the decomposition of an arbitrary grid cell into components that are due to anthropogenic and biogenic contributions.

Focusing here on our biogenic modeling contributions, we plan to use physics-constrained machine learning to inform the transformer architecture's characterization of biogenic CH_4 fluxes. Specifically, we are incorporating a bio-geophysical CH_4 production model into a component of ecosystem embedding transformer objective function.

Ultimately, our combined source-specific bottom-up modeling approach will augment top-down monitoring efforts by allowing researchers to ask questions about consensus between measurements and models and, most critically, to understand how terrestrial CH_4 production is changing as a function of natural and human caused activities—a distinction that is central to managing and mitigating climate change.

Implementing and Benchmarking an Agricultural Methane Emissions Model in E3SM

Kendalynn Morris,[1] Abigail Synder,[1] Eva Sinha,[2] Sha Feng[2]

[1]Joint Global Change Research Institute, [2]Pacific Northwest National Laboratory

Focal Areas

A major knowledge gap in our understanding of global carbon cycling is the current and future role of croplands in both production and consumption of atmospheric methane (CH_4). One of the most powerful tools available for assembling and testing our knowledge of CH_4 flux are Earth system models such as E3SM. The current E3SM Land Model (ELM) does not consider any managed ecosystem CH_4 flux. Therefore, expanding ELM's capabilities, currently limited to wetlands, to include croplands will leverage emerging datasets from recent syntheses (Guo et al. 2023), eddy covariance networks (Delwiche et al. 2021), and top-down CH_4 flux estimates (Hannun et al. 2020) while building on the expertise already established in ELM.

Science or Technological Challenge

The challenge of this expansion is three-fold: (1) processes currently parameterized for wetland CH_4 flux will not directly translate to managed upland systems (Riley et al. 2011); (2) while data availability is improving rapidly, CH_4 flux and corresponding biotic and abiotic metadata from croplands are not as extensive as for wetland ecosystems; and (3) croplands are dynamically managed, requiring an understanding of the economic context that drives crop production and feeds back into CH_4 flux. We propose that machine learning (ML) approaches deployed in combination with domain expertise and additional DOE-supported research products, can bridge these challenges and support the development and testing of a process-based, interpretable model.

Rationale

Most process-based ecosystem CH_4 emission models are oriented towards wetland ecosystems that are large natural producers of CH_4. However, the global CH_4 emissions from croplands, primarily from rice cultivation, are estimated to be 8% of global anthropogenic CH_4 emissions (Saunois et al. 2020). Upland ecosystems can also be CH_4 sinks, and active management of croplands, such as periodic drainage of rice paddies (Runkle et al. 2019) and no-till agriculture (Ussiri et al. 2009), have the potential to offset CH_4 release. Furthermore, global change is increasing both the magnitude of cropland sink potential and the frequency of intense rainfall events that could shift these systems to CH_4 sources. This combined with the dynamic potential of various agricultural management practices and the strong radiative forcing effect of CH_4 makes incorporating these agroecosystems in Earth system models increasingly important. However, the spatiotemporal variability in CH_4 flux in cropland ecosystems, limitations of data availability on that flux, and the managed-lands aspect of these ecosystems all represent distinct challenges to this advancement.

Narrative

Our goal is to implement an AI-informed, process-based cropland CH_4 emission module within E3SM. The following steps outline our approach in broad strokes:

Step 1. Even with domain expertise, it is not immediately obvious what processes or parametrizations should be prioritized to update process-based models from wetland to cropland, particularly given the more limited data available for cropland. As a first step, we propose to utilize exploratory, unsupervised ML to identify reduced-form patterns explaining a meaningful proportion of variance in the spatiotemporal CH_4 datasets for different areas, including: (1) observed wetland CH_4 data, (2) simulated wetland CH_4 data, (3)(albeit more limited) observed cropland CH_4 data, and (4) cropland CH_4 data simulated with ELM's wetland CH_4 model. These datasets would include variables such as measured (or modeled) CH_4 concentration, meteorological data, temperature, humidity, and soil temperature. Self-Organizing Maps (SOMs) are a promising method to guide exploration of existing data sources to prioritize aspects for updating. By training SOMs on these datasets, a lower-dimensional representation (or generated map) of each will be produced (Nourani et al. 2013). Each generated map is an extraction of complex patterns characteristic to the training data, and the direct comparison of the resulting patterns will allow us to explore differences in cropland versus wetland CH_4 processes, thereby guiding development efforts in the next step.

More specifically, each generated map can treat the three nontraining datasets as novel input data to be classified, essentially displaying the nontraining dataset in the space of the training data's distribution in a visually digestible way. The resulting figures can be used to identify discrepancies among these datasets and guide approaches to adapt, update, and/or reparameterize the well-established wetland-CH_4 processes for cropland. Some essential additions are already known, for instance, ELM currently does not model rice, the dominant crop when considering agricultural CH_4 sources. However, interpreting these maps and their discrepancies fundamentally requires domain expertise because it is an exploratory exercise.

This results in expertise-guided hypotheses of updates to the process-based wetland model for use in cropland that can be made and interpreted iteratively.

Step 2. After this exploratory phase, the revised cropland-CH_4 module will undergo quantitative assessment. This validation will come from comparing simulated ELM cropland CH_4 versus FLUXNET-CH4 data. Here we will implement classical analysis of error between simulated and observed cropland CH_4 values as recommended by International Land Model Benchmarking (Collier et al. 2018). Additionally, we will use ML approaches to characterize multidimensional spatiotemporal error (Tebaldi et al. 2021) to highlight areas of improvement missed by classical multimetric approaches.

Step 3. As knowledge gaps are identified via Steps 1 and 2, literature review and synthesis using emerging data on upland CH_4 flux (Guo et al. 2023) will be used to fill these gaps, when possible, following statistically rigorous meta-analysis techniques (Morris et al. 2022). Additionally, ongoing work as part of DOE's COMPASS project will provide valuable syntheses of upland CH_4 sink-to-source transition points.

Step 4. We hypothesize that land management practices are crucially important to capturing variability in cropland CH_4 flux. Therefore, the final component of this proposed research is to incorporate land management practices that can impact cropland CH_4 emissions and to use the updated model to quantify CH_4 mitigation that can be achieved in the future under various climate change scenarios. If our hypothesis is correct, expanding the current management options of ELM's cropland module to include soil drainage and aeration will be essential. One possibility that leverages additional expertise would be incorporation of the land-use and agricultural technology distributions from the Global Change Assessment Model (GCAM), which is now actively coupled into E3SM, opening exciting simulation possibilities in this area. GCAM is an integrated assessment model that takes into consideration the land-energy-human-climate system. Such models are the best tools available for assessing various global C management scenarios. An aspect of the cropland CH_4 module would then be the ability to reflect different, albeit estimated, adaptation rates of conservation agricultural practices under various policy scenarios.

References

Collier, N., et al. 2018. "The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation," *Journal of Advances in Modeling Earth Systems* **10**(11), 2731–54.

Delwiche, K. B., et al. 2021. "FLUXNET-CH_4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–89.

Guo, J., et al. 2023. "Global Climate Change Increases Terrestrial Soil CH_4 Emissions," *Global Biogeochemical Cycles* **37**(1). DOI:10.1029/2021gb007255.

Hannun, R. A., et al. 2020. "Spatial Heterogeneity in CO_2, CH_4, and Energy Fluxes: Insights from Airborne Eddy Covariance Measurements over the Mid-Atlantic Region," *Environmental Research Letters* **15**, 035008.

Morris, K. A., et al. 2022. "Soil Respiration Response to Simulated Precipitation Change Depends on Ecosystem Type and Study Duration," *Journal of Geophysical Research: Biogeosciences* **127**, e2022JG006887. DOI:10.1029/2022jg006887.

Nourani, V., et al. 2013. "Using Self-Organizing Maps and Wavelet Transforms for Space–Time Pre-Processing of Satellite Precipitation and Runoff Data in Neural Network Based Rainfall–Runoff Modeling," *Journal of Hydrology* **476**, 228–43.

Riley, W. J., et al. 2011. "Barriers to Predicting Changes in Global Terrestrial Methane Fluxes: Analyses Using CLM4Me, a Methane Biogeochemistry Model Integrated in CESM," *Biogeosciences* **8**(7), 1925–953.

Runkle, B. R. K., et al. 2019. "Methane Emission Reductions from the Alternate Wetting and Drying of Rice Fields Detected Using the Eddy Covariance Method." *Environmental Science & Technology* **53**(2), 671–81.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12**(3),1561–1623.

Tebaldi, C., et al. 2021. *Machine Learning for a-Posteriori Model-Observed Data Fusion to Enhance Predictive Value of ESM Output*, AI4ESP1131. Joint Global Change Research Institute, Pacific Northwest National Laboratory. DOI:10.2172/1769740.

Ussiri, D. A. N., et al. 2009. "Nitrous Oxide and Methane Emissions from Long-Term Tillage Under a Continuous Corn Cropping System in Ohio," *Soil and Tillage Research* **104**(2), 247–55.

AI4 Plant Trait-Based Wetland CH_4 Predictions

Avni Malhotra,[1] Tiia Määttä,[2] Etienne Fluet-Chouinard,[1] Housen Chu,[3] Gavin McNicol,[4] Kyle Delwiche[5]

[1]Pacific Northwest National Laboratory, [2]University of Zurich, [3]Lawrence Berkeley National Laboratory, [4]University of Illinois–Chicago, [5]University of California–Berkeley

Focal Areas

- Approaches that support the transfer of mechanistic knowledge gained in the laboratory to make predictions in the field, and vice versa.
- Key uncertainties and knowledge gaps in CH_4 where new AI technology can advance plant-trait based predictive understanding of the wetland methane cycle.
- The importance of high-potential datasets (FLUXNET-CH4; COSORE; plant/root trait databases; network data and experiments such as NEON, COMPASS, and SPRUCE) and how the combination of data across spatial or temporal scales or scientific domains may lead to new scientific insights, either within or across fields.

Science or Technological Challenge

Predicting highly variable wetland CH_4. Wetlands are the largest natural source of CH_4 to the atmosphere and remain a key uncertainty in the global CH_4 budget, emitting between 100–180 Tg CH_4 yr^{-1} (Saunois et al. 2020). Wetlands also face unique pressures (drainage, salinization, etc.) from human land uses (Fluet-Chouinard et al. 2023) and climate change (Peng et al. 2022), often driving these systems into disequilibrium (Camill and Clark 1998). Uncertainty in wetland CH_4 emissions is partly due to the dynamic nature of wetland biogeochemistry and hydrology, as well as processes involved in CH_4 flux (methanogenesis, methanotrophy, gas transport, etc.). The variability of wetland ecosystem structure and function is hypothesized to further increase with increasing environmental stressors (Malhotra and Roulet 2015) and expected to further hinder CH_4 predictions and scaling.

Plants are integrators of the high spatiotemporal variability in wetland ecosystems, responding to and influencing microbial structure and function, soil moisture, nutrient status, etc. Thus, fine-scale (~1 m^2/hourly to 1 km^2/season) heterogeneity in plant properties (hereafter, traits) is often closely related to wetland CH_4 flux variability (Waddington et al. 1996; Lai et al. 2014; Goud et al. 2017; Knox et al. 2021), and plant trait incorporation into empirical and predictive models of CH_4 could help reduce uncertainties from fine-scale variability. Advances in high-potential CH_4 flux databases (Knox et al. 2019; Delwiche et al. 2021), plant trait databases (Iversen et al. 2017; Kattge et al. 2019), wetland CH_4 modeling (Salmon 2022), and deep neural network technologies (Chen et al. 2018; Reichstein et al. 2019) combined with process knowledge derived from controlled laboratory and manipulative field experiments can help refine our understanding and predictions of wetland CH_4.

Rationale

Above and belowground plant traits to improve CH_4 predictions. Advances in incorporating a mechanistic and scalable understanding of how plant traits influence wetland CH_4 emissions have been hindered by several research gaps. (1) We lack a synthetic view of which plant traits most affect CH_4 processes and can be best used as predictors. Knowledge gaps particularly remain around the mechanistic links between root traits and CH_4 fluxes (e.g., root biomass, rooting depth, exudation, aerenchyma size; Sutton-Grier and Megonigal 2011). (2) While chamber-based CH_4 measurements are often coupled with plant trait information, quantification of wetland plant traits at the footprint scale of CH_4 eddy covariance towers is usually difficult. (3) Also lacking are frameworks to connect relatively-easy-to-measure and remotely sensible aboveground with belowground traits. (4) Until recently, high-potential validation datasets were unavailable on CH_4 flux and plant traits, particularly root traits, that would allow for scaling mechanistic information from lab and field studies to site and regional scales.

Narrative

AI-enabled mechanistic linkages between plant traits and wetland CH_4. We propose to incorporate lab/field-scale mechanistic understanding of plant trait drivers of CH_4 into site and regional scales using a combination of lab and field studies, data syntheses, and deep neural network modeling. Our approach is broadly divided into two steps:

1. Developing and synthesizing mechanistic frameworks from new lab studies, and existing gradients and experiments. Controlled laboratory studies, such as wetland soil incubation experiments with isotopically labeled root material to trace the fate of root-carbon in CH_4 emissions, will be used as one tool to generate functions of CH_4 response to trait variability. We will also synthesize plant trait and CH_4 data from natural gradient studies and databases, such as NEON and FRED (Iversen et al. 2017), and from existing manipulative experiments to provide a gradient of plant trait values. For example, SPRUCE (Hanson et al. 2017) provides root trait and CH_4 flux gradients across a peatland warming study (Hanson et al. 2020; Malhotra et al. 2020) and COMPASS sites provide elevational gradients across coastal wetlands. We will also partner with ongoing experimental data synthesis efforts such as the DeepSOIL2100. Through these lab and synthetic studies, we will develop specific model structures and parameters on trait-CH_4 links for our AI predictions in (2).

2. Predictive modeling of the mechanistic links between plant traits and wetland CH_4. We will test mechanistic model structures developed in (1) linking CH_4-relevant plant traits and CH_4 processes as frameworks for hybrid AI methods such as neural ordinary differential equations (neural ODE; Chen et al. 2018). Such hybrid approaches allow learning process parameters, latent variables, and functional relationships across a number of hypothesized structural constraints and complexity. Input data for these models will originate from experiments and syntheses highlighted in (1). High-potential datasets such as the FLUXNET-CH4 and COSORE databases (Bond-Lamberty et al. 2020; Delwiche et al. 2021) would serve as the key validation datasets. In particular, combining chamber and eddy covariance tower measurements from the same sites will allow us to test trait-CH_4 linkages in a neural ODE across spatial scales to evaluate the generalizability of the learned parameters and functional relationships. Through the integration of lab, field, and synthetic data into a hybrid modeling approach, this project will allow us to identify key mechanistic constraints and sources of uncertainty of the relationship between plant traits and CH_4 emissions in wetlands.

References

Bond-Lamberty, B., et al. 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology* **26**(12), 7268–283.

Camill, P., and J. S. Clark. 1998. "Climate Change Disequilibrium of Boreal Permafrost Peatlands Caused by Local Processes," *American Naturalist* **151**(3), 207–22.

Chen, R. T. Q., et al. 2018."Neural Ordinary Differential Equations," *arXiv* 1806, 07366 [cs.LG].

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–689.

Fluet-Chouinard, E., et al. 2023. "Extensive Global Wetland Loss Over the Past Three Centuries," *Nature* **614**, 281–86.

Goud, E. M., et al. 2017. "Predicting Peatland Carbon Fluxes from Non-Destructive Plant Traits," *Functional Ecology* **31**(9), 1824–833.

Hanson, P. J., et al. 2017. "Attaining Whole-Ecosystem Warming Using Air and Deep-Soil Heating Methods with an Elevated CO_2 Atmosphere," *Biogeosciences* **14**(4), 861–83.

Hanson, P. J., et al. 2020. "Rapid Net Carbon Loss from a Whole-Ecosystem Warmed Peatland," *AGU Advances* **1**(3), e2020AV000163. DOI:10.1029/2020av000163.

Iversen, C. M., et al. 2017. "A Global Fine-Root Ecology Database to Address Below-Ground Challenges in Plant Ecology," *New Phytologist* **215**(1), 15–26.

Kattge, J., et al. 2019. "The Global Database of Plant Traits: TRY Version 5.0," *Geophysical Research Abstracts* **21**, EGU2019-18965.

Knox, S. H., et al. 2019. "FLUXNET-CH4 Synthesis Activity: Objectives, Observations, and Future Directions," *Bulletin of American Meteorological Society* **100**(12), 2607–632. DOI:10.1175/BAMS-D-18-0268.1.

Knox, S. H., et al. 2021. "Identifying Dominant Environmental Predictors of Freshwater Wetland Methane Fluxes Across Diurnal to Seasonal Time Scales," *Global Change Biology* **27**(15), 3582–604.

Lai, D. Y. F., et al. 2014. "Spatial and Temporal Variations of Methane Flux Measured by Autochambers in a Temperate Ombrotrophic Peatland," *Journal of Geophysical Research: Biogeosciences* **119**(5), 864–80.

Malhotra, A., et al. 2020. "Peatland Warming Strongly Increases Fine-Root Growth," *PNAS* **117**(30), 17627–634.

Malhotra, A., and N. T. Roulet. 2015. "Environmental Correlates of Peatland Carbon Fluxes in a Thawing Landscape: Do Transitional Thaw Stages Matter?" *Biogeosciences* **12**(10), 3119–130.

Peng, et al. 2022. "Wetland Emission and Atmospheric Sink Changes Explain Methane Growth in 2020," *Nature* **612**, 477–82.

Reichstein, M., et al. 2019. "Deep Learning and Process Understanding for Data-Driven Earth System Science," *Nature* **566**, 195–204.

Salmon, E., 2022. "Assessing Methane Emissions for Northern Peatlands in ORCHIDEE-PEAT Revision 7020," *Geoscientific Model Development* **15**(7), 2813–838.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12**(3), 1561–1623. DOI:10.5194/essd-12-1561-2020.

Sutton-Grier, A. E., and J. P. Megonigal. 2011. "Plant Species Traits Regulate Methane Production in Freshwater Wetland Soils," *Soil Biology and Biochemistry* **43**(2), 413–20.

Waddington, J. M., et al. 1996. "Water Table Control of CH_4 Emission Enhancement by Vascular Plants in Boreal Peatlands," *Journal of Geophysical Research* **101**(D17), 22775–785.

Expanding Eddy Covariance Measurements from Tropical Wetland Methane Emissions to Improve AI-Aided Emissions Upscaling

Kyle Delwiche,[1] Rob Jackson,[2] Sara Knox,[3] Avni Malhotra,[4] Etienne Fluet-Chouinard,[4] Alison Hoyt,[2] Zutao Ouyang,[2] Gavin McNicol,[5] Trevor Keenan,[1]

[1]University of California–Berkeley, [2]Stanford University, [3]The University of British Columbia, [4]Pacific Northwest National Laboratory, [5]University of Illinois–Chicago

Focal Areas

Expanding the network of methane eddy covariance measurements in tropical wetland ecosystems by facilitating new data collection and adding existing data into an updated version of the FLUXNET-CH4 dataset. Combining this new network and AI/ML-based models with recent advances in mapping tropical inundation and wetland types to improve process-based emission models and enhance their agreement with top-down estimates of tropical methane emissions, such as satellite-based instruments.

Science or Technological Challenge

Global atmospheric methane (CH_4) concentrations are rising at an accelerating rate, yet uncertainties around major terrestrial and aquatic CH_4 sources currently prevent global CH_4 budget closure. Tropical latitudes account for roughly 68% of global emissions, and most tropical emissions come from wetlands (Saunois et al. 2020). Despite their importance to global methane emissions, current tropical wetland methane emissions and projected changes due to climate change are poorly understood. This uncertainty is due to multiple factors, including the paucity of field-based measurements of methane emissions from the tropics (Delwiche et al. 2021), a lack of process-based understanding of tropical methane emissions (Parker et al. 2018), and uncertainties surrounding inundation mapping both currently and under future hydrological change (Gerlein-Safdi et al. 2021; Padney et al. 2021). Accurately upscaling tropical wetland methane emissions will therefore require more field-based measurements (particularly from eddy covariance towers and flux chambers), AI-fueled advances in upscaling techniques, and improved modeling of hydrological and ecophysiological factors governing methane release.

Rationale

Eddy covariance systems are able to measure methane emissions at high temporal resolution over landscape scales, yet these systems require significant maintenance and are therefore less utilized in challenging tropical ecosystems. Furthermore, eddy covariance methane data that are currently being collected in tropical ecosystems are not all included in FLUXNET-CH4, the global compilation of methane flux data (Delwiche et al. 2021), for a variety of practical, technical, and cultural reasons. Increasing the availability of these valuable datasets requires a systematic effort to work with local site teams to QA/QC, partition, and gap-fill eddy covariance methane and CO_2 data for inclusion in an updated version of FLUXNET-CH4. Methodological advances enabling the combination of EC data with other high-potential datasets such as chamber flux databases (Bond-Lamberty et al. 2020), would improve spatial representation and prediction. Our recent work using AI-aided models to upscale global wetland CH_4 fluxes has found strong model divergence in humid tropical rainforest regions, highlighting the need for more tropical data. These new gridded products from upscaling global wetland CH_4 fluxes (UpCH4; McNicol et al. In preparation) and monsoon Asia paddy-rice CH_4 fluxes (RiceCH4; Ouyang et al. 2023) both have insufficient training sites in the tropics. For UpCH4, this leads to large differences between upscaled products and state-of-the-art process and inversion-based models. However, better model convergence in ecosystems with more training data (temperate and boreal ecosystems) demonstrates the potential of AI-driven upscaled products as long as sufficient training data exist.

Two other critical needs for improving estimates of tropical wetland methane emissions are better maps classifying tropical wetlands and refined inundation maps. While inundation alone is insufficient to explain tropical CH_4 fluxes, which also vary with nutrient dynamics, vegetation, and carbon inputs, inaccurate tropical hydrology can exacerbate mismatches between process-based model estimates of methane emissions and satellite-based measurements from GOSAT or TROPOMI (Parker et al. 2018; Pandey et al. 2021). For example, by including inundation dynamics in wetland maps and linking these to methane emissions models, recent work by Gerlein-Safdi et al. (2021) resulted in improvements to the predicted seasonality and interannual variability of methane emissions. Thus, more work is needed to develop AI tools to detect inundation conditions over time with L-band radar capable of making measurements through cloud or vegetation cover, such as CYGNSS (Zeiger et al. 2022) and NISAR.

Narrative

To expand the amount of eddy covariance data available from tropical wetland ecosystems, we will develop partnerships with research groups currently making tropical eddy covariance measurements. In some cases, existing flux towers currently not measuring methane flux will be retrofitted to include methane sensors. We will identify existing methane eddy flux datasets and work with site PIs to QA/QC data in preparation for inclusion in FLUXNET-CH4 Version 2.0. We will enhance tropical flux science by holding pantropical workshops on eddy covariance data processing to foster regional partnerships for technical guidance and knowledge transfer. We will build on the relationships developed during workshops and technical assistance to support the establishment of new flux towers in under-studied ecosystems, particularly in Africa and South America.

New flux tower sites will be located along hydrological gradients to address the seasonality of local flooding conditions, and gradient data will be paired with improvements in inundation mapping to enhance AI-driven upscaling of tropical methane emissions. In addition to establishing new flux tower sites, new chamber measurements will be taken across sites with eddy covariance towers to aid in chamber/tower data comparisons. This will allow us to develop the workflow to reconcile chamber- and EC-based flux measurements to gain better spatial representation, building upon ongoing chamber tower comparisons (Määttä et al. In preparation).

This work to expand tropical wetland methane flux measurements will directly support existing efforts to upscale FLUXNET-CH4 data using AI/ML models. Currently, efforts are hampered by lack of data in the tropics, so this proposed expansion in tropical datasets will greatly improve our ability to estimate tropical wetland contributions to the global methane budget, as well as projected future changes in emissions under climate change. The upscaling work will be supported by improved seasonal inundation maps in tropical ecosystems, as well as the incorporation of new metadata and controlled vocabularies required for tropical wetland ecosystems (e.g., updating plant functional types to include tropical systems).

References

Bond-Lamberty, B., et al., 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology*, **26**(12), 7268–283. DOI:10.1111/gcb.15353.

Delwiche, K. B., et al., 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–89. DOI:10.5194/essd-2020-307.

Määttä T. "Spatial Heterogeneity Dictates Coherence Between Eddy Covariance- and Chamber-Based CH_4 Measurements Across Multiple Wetland Sites." In preparation.

McNicol, G., et al. "UpCH4: A Global Freshwater Wetland Methane Emissions Product for 2001–2018 From Upscaled Eddy Covariance Fluxes." In preparation.

Ouyang, Z., et al. 2023. "Paddy Rice Methane Emissions Across Monsoon Asia," *Remote Sensing of Environment* **284**, 113335. DOI:10.1016/j.rse.2022.113335.

Pandey, S., et al., 2021. "Using Satellite Data to Identify the Methane Emission Controls of South Sudan's Wetlands," *Biogeosciences* **18**(2), 557–72. DOI:5194/bg-18-557-2021.

Parker, R. J., et al., 2018. "Evaluating Year-to-Year Anomalies in Tropical Wetland Methane Emissions Using Satellite CH_4 Observations," *Remote Sensing of Environment* **211**, 261–75. DOI:10.1016/j.rse.2018.02.011.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12**(3), 1561–1623. DOI:10.5194/essd-12-1561-2020.

Zeiger, P., et al. 2022. "Introducing the Global Mapping of Flood Dynamics Using GNSS-Reflectometry and the CYGNSS Mission," *ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Sciences* **V-3-2022**, 93–100. DOI:10.5194/isprs-annals-v-3-2022-93-2022.

A Hybrid Approach to Improve Earth System Model Predictions of CH_4 Emissions from Northern Peatlands

Dan Ricciuto,[1] Dan Lu,[1] Dali Wang,[1] Xiaoying Shi,[1] Xiaofeng Xu,[2] Melanie Mayes,[1] and Paul Hanson[1]

[1]Oak Ridge National Laboratory, [2]San Diego State University

Focal Areas

- Application of AI-based surrogate modeling approaches.
- Earth system modeling.
- Model optimization and calibration.

Scientific or Technological Challenge

The accurate prediction of CH_4 emissions from northern peatland systems is crucial for understanding the role of terrestrial ecosystems and their feedbacks to the global carbon cycle. However, simulating CH_4 emissions is challenging as peatlands are complex ecosystems that contain a range of interacting processes, including vegetation productivity and litter inputs, hydrology, microbial decomposition, and different pathways for the production and consumption of CH_4. In particular, those interactions are nonlinear and vary strongly across space and over time. Meanwhile, global environmental changes (i.e., climate change and elevated CO_2) add complexity to CH_4 modeling. Capturing all variations and future climate scenarios into reliable peatland models of CH_4 fluxes is a challenge that needs to be addressed.

Rationale

There is a clear need for improved models to accurately predict CH_4 fluxes given the significant impact that these emissions can have on the global climate. Although there are a number of CH_4 observation sites, data from manipulative treatments are limited. The Spruce and Peatland Responses Under Changing Environments (SPRUCE) experiment introduced whole ecosystem warming and elevated CO_2 treatments into an ombrotrophic bog in northern Minnesota, and initial results have indicated strong increases in CO_2 and CH_4 fluxes with warming (Hanson et al. 2020). The warming response of CH_4 flux is highly sensitive to water table position, as evidenced by a 2021 drought that strongly reduced CH_4 emissions. It is currently unclear whether the results from SPRUCE are representative of other wetland systems, because the SPRUCE treatments take the system far beyond what can be determined from the range of natural variability at other sites (Helbig et al. 2022). Additionally, running model experiments with Earth Land Model (ELM)-SPRUCE to cover large spatial domains or parametric uncertainty is computationally expensive. Our proposed approach uses AI methods to extend our peatland model, ELM-SPRUCE, to other wetlands using neural network-based surrogate modeling approaches. Our approach will provide high-resolution maps of CH_4 fluxes and their uncertainties over northern peatlands in North America under historical and future climate conditions.

Narrative

Aim 1: Develop a hybrid modular vegetation and hydrology framework. In this framework, we could replace expensive model components in ELM-SPRUCE with surrogate AI-based representations. Simulating gross primary production and canopy-scale fluxes requires an hourly or smaller time step and a computationally demanding solution. We may use model output to train a neural network representation that can predict these fluxes at any desired temporal resolution as a function of meteorological drivers and two key model state variables: soil moisture and leaf area index. This module may then be replaced with this surrogate model that is much faster to evaluate. A second surrogate model representation will be developed for predicting soil moisture, temperature, and water table position. Predicting these variables requires knowledge about past states and drivers. A recurrent neural network is likely a good choice for making these predictions. We could use an interpretable LSTM (iLSTM; Lu et al. 2022a) to incorporate these memory effects and to provide physical insights about key drivers. The two surrogate models may be connected to the physically based ELM model of vegetation allocation and turnover to predict leaf area index and litterfall. This submodel is computationally inexpensive and may be simulated at daily or greater time steps. This combined modeling system would provide estimates of litterfall, nutrient demand, and soil conditions for the decomposition model (Aim 2).

Aim 2: Develop physical and surrogate representations of decomposition and CH4 models. We can use the standalone version of the Microbe model (Xu et al. 2015; Riccuto et al. 2021) that is currently connected to ELM-SPRUCE to predict CH_4 production through hydrogenotrophic and acetoclastic methanogenesis, oxidation, and flux to the atmosphere through plant transport, diffusion, and ebullition. The associated decomposition model will also estimate nutrient mineralization that can be coupled to the vegetation allocation model (Aim 1). It would be computationally feasible to perform a large ensemble of Microbe model simulations, capturing the impacts under a wide range of temperature and moisture conditions, soil carbon distributions, litter inputs, and parametric uncertainty on CH_4 fluxes predicted by the model. This large ensemble could be used to train a surrogate model. It is unclear which machine learning or AI method will work best for this surrogate model, and we might explore multiple methods considering both spatial and temporal properties of the simulated fields.

Aim 3: Model calibration and regional simulation. We can calibrate the hybrid model in Aim 1 coupled with the surrogate Microbe model in Aim 2 to obtain posterior parameter distributions for SPRUCE and AmeriFlux sites given CH_4 flux observations. We can use Markov Chain Monte Carlo to obtain these distributions and may also explore using invertible neural networks (Lu et al. 2022b) to improve the efficiency of the calibration process. Scaling the results to boreal North American peatlands can be done at high resolution using the Peat-ML product (Melton et al. 2022) or other similar products to define peatland areas and initial peatland carbon stocks. Downscaled DAYMET data is available for historical simulations, and we could also perform future simulations using downscaled outputs from CMIP6 Earth system models (Rastogi et al. 2021). Historical and future projections of CH_4 fluxes and their uncertainties could be made available to the broader community for further analysis.

References

Hanson, P. J., et al. 2020. “Rapid Net Carbon Loss from a Whole-Ecosystem Warmed Peatland,” *AGU Advances* **1**(3), e2020AV000163. DOI:10.1029/2020AV000163.

Helbig, M., et al. 2022. “Warming Response of Peatland CO_2 Sink is Sensitive to Seasonality in Warming Trends,” *Nature Climate Change* **12**(8), 743–49. DOI:10.1038/s41558-022- 01428-z.

Lu, D., et al. 2022a. “An Interpretable Machine Learning Model for Advancing Terrestrial Ecosystem Predictions,” ICLR Conference AI for Earth and Space Sciences Topic.

Lu, D., et al. 2022b. “Invertible Neural Networks for E3SM Land Model Calibration and Simulation,” ICLR Conference AI for Earth and Space Sciences Topic.

Melton, J. R., et al. 2022. “A Map of Global Peatland Extent Created Using Machine Learning (Peat-ML),” *Geoscientific Model Development* **15**(12), 4709–38. DOI:10.5194/gmd-15-4709-2022.

Rastogi, D., et al. 2022. “How May the Choice of Downscaling Techniques and Meteorological Reference Observations Affect Future Hydroclimate Projections?” *Earth’s Future* **10**(8), e2022EF002734. DOI:10.1029/2022EF002734.

Ricciuto, D. M., et al. 2021. “An Integrative Model for Soil Biogeochemistry and Methane Processes: I. Model Structure and Sensitivity Analysis,” *Journal of Geophysical Research: Biogeosciences* **126**(8), e2019JG005468. DOI:10.1029/2019JG005468.

Xu, X., et al. 2015. “A Microbial Functional Group-Based Module for Simulating Methane Production and Consumption: Application to an Incubated Permafrost Soil,” *Journal of Geophysical Research: Biogeosciences* **120**(7), 1315–33. DOI:10.1002/2015JG002935.

The Potential for Artificial Intelligence to Inform Pore-Scale Patterns of Methane Production, Release, and Consumption Using Imaging, Real-Time Flux Measurements, and Microbial Modeling

Melanie Mayes,[1] Dan Lu,[1] Tamas Varga,[2] Xiaofeng Xu,[3] Jeff Warren,[1] Kristin Boye,[4] Vincent Noël,[4] Alex Johs,[1] Elizabeth Herndon,[1] Dan Riccuito,[1] Paul Hanson[1]

[1]Oak Ridge National Laboratory; [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory; [3]San Diego State University; [4]SLAC National Accelerator Laboratory

Focal Areas

- New methodology/technology
- High-resolution automated datasets
- Lab to field

Scientific or Technological Challenge

The production, release, and consumption of methane are pore-scale processes with global impact. We currently lack the capability to observe and understand pore-scale controls over the methane cycle, such as the dynamic role of soil moisture and oxygen content, the configuration of complex pore structures like microsites in soils, and the interplay of different microbial functional groups (e.g., acetogens, acetoclastic and hydrogenotrophic methanogens, methanotrophs). Consequently, models at every scale lack the ability to predict net soil methane emissions at the ecosystem level given the substantial heterogeneities in soil aggregate size, shape, physico-chemical properties, and temporal dynamics.

Rationale

We need a better understanding of how methane is produced within and released from soils, but there are numerous complexities that have thus far inhibited understanding and prediction of gross and net methane fluxes from soils. Methane production (methanogenesis) and consumption (methanotrophy) occur in soil pores, where heterogeneities in pore size and configuration control O_2 levels and moisture content (Silver et al. 1999). Methanogenesis is restricted to anaerobic conditions, which can develop rapidly in response to increases in soil moisture and can also persist in microsites long after soil macropores have drained and reoxygenated. Changes in moisture can differentially affect substrate supply for acetoclastic methanogens (solute substrate) and hydrogenotrophic methanogens (gas substrate; Sihi et al. 2021). Methanotrophs tend to thrive in aerobic conditions, but aerobic conditions can also persist inside microsites, fueling the consumption of methane even under anaerobic conditions. The advent of technologies for simultaneous measurement of methane and carbon dioxide in automated flux chambers has greatly increased the ability to observe high-resolution temporal dynamics [O'Connell et al. 2018, 2022; Bond-Lamberty et al. 2020 (CO_2 only)]. These kinds of surface soil measurements can constrain net methane fluxes, but the complexity and spatiotemporal dynamics inside soils remain a mystery that inhibits broader methane predictability. The same principles and techniques needed for methane can also be applied to other redox-sensitive processes like nitrous oxide emissions and metal redox transformations.

Narrative

Aim 1: Collect imaging and geochemical data under different moisture and O_2 scenarios. The evolution, transport, and release dynamics of soil methane can be measured in aggregate-scale microenvironments. While measurements in cores may be limited to net fluxes, the information content can be greatly improved by high temporal resolution imaging and spectroscopy. Imaging technologies like CT-scanning and neutron tomography could be used to map soil moisture and gas content in structured materials such as soil aggregates and cores. Neutron imaging could be used to assess methane gas bubble transport rates and resistances within porous media. Bubbles (void space) would be visualized as lighter spots by imaging against a darker background at scales of 25–150 μm. X-ray CT-scanning could be accomplished on aggregates or mini-cores (diameter ~1–2 cm) to provide 20 μm resolution or resolution of a few μm with micro-CT. Soil cores of 5–10 cm in diameter can be used to better understand processes in macropores and provide correlations with bulk properties, such as gas flux, water saturation, and organic C content. Porosity information (e.g., pore volume fraction, pore size, and pore connectivity) from imaging can be related to methane storage and release. 3D imaging can be used to reconstruct or track a bubble as it migrates through time, at scales of sub-µm to mm. Imaging may also identify active microorganisms indicative of "hot spots" of activity. At the smallest scales, synchrotron spectroscopy can identify metal redox species in the solid phase, which can be key indicators of the redox environment and can alter the direction and outcomes of redox reactions (e.g., iron and sulfate reduction can inhibit methanogenesis; Bear et al. 2021).

Aim 2: Configure AI models to match imaging, geochemical, and methane flux data. Convolutional neural networks (CNN) can be used to learn the relationships between soil structure and methane fluxes, and soil moisture and O_2 content (Liu et al. 2022). AI can generate new images to fill the data gaps using generative models, such as GAN for normalizing flow and diffusion models. There are also ML models for feature extraction, segmentation, and clustering; these can be used to guide and optimize image segmentation during CT or neutron scanning (e.g., Venkatakrishnan et al. 2021). We can use a regionalized CNN model for image segmentation to extract interesting features and learn the pore volume in the soil core from the image, and also learn the relationships between pore volume and methane flux, and moisture and O_2 content. For time series data, we can use a long short-term memory network. If the data is an image, we can use CNN. If the data is like a network containing both spatial and temporal information, we can use a graph neural network.

Aim 3: Allow AI models to constrain model processes and provide parameters for different scenarios. We can use AI to test and parameterize existing models (Xu et al. 2015; Sihi et al. 2021) that contain key mechanisms, such as acetoclastic and hydrogenotrophic methanogens, methanotrophs, acetogens, dissolved organic carbon supply, sulfate concentrations, O_2 concentrations, and pH changes. Other measured constraints, such as Eh and other alternative electron acceptors, could aid convergence of the AI and microbial functional models. ML models can facilitate process-based model simulations by building a fast-to-evaluate surrogate model to reduce computational cost of the process-based model to facilitate parameter estimation or uncertainty quantification. Invertible neural networks can be used to build a surrogate model and estimate the model parameters at the same time as the process-based models, thereby permitting convergence between process- and ML-models.

Aim 4: Match the trained AI model to complex field- and lab-scale data to enable site-level predictions and beyond. Scaling up, the trained AI model could be used to match existing automated chamber data at lab and field scales (O'Connell et al. 2018, 2021; Sihi et al. 2021; new datastreams at the SPRUCE experiment). When connected with time series data (e.g., soil moisture, O_2, methane fluxes, Eh, etc.), interpretable ML models can help explain the importance of various drivers and their contributions to bulk methane flux predictions from the field. We can quantify the uncertainty of both the ML- and process-based model; analyze the contribution of prediction uncertainty from the model structure, model parameter, and data; and use this uncertainty analysis to guide data collection and further improve model development. Finally, ML can assist with multi-scale modeling to extrapolate learning to sites lacking imaging and pore-scale data. This project would provide new insights into key microsite dynamics and improve prediction and controls over net methane fluxes in soils.

References

Bear, S. E., et al. 2021. "Beyond the Usual Suspects: Methanogenic Communities in Eastern North American Peatlands are also Influenced by Nickel and Copper," *FEMS Microbiology Letters* **368**(21–24).

Bond-Lamberty, B., et al. 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology* **26**(12), 726883.

Liu, S., et al. 2022. "Improving Net Ecosystem CO_2 Flux Prediction Using Memory-Based Interpretable Machine Learning," *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1111–19.

O'Connell, C. S., et al. 2018. "Drought Drives Rapid Shifts in Tropical Rainforest Soil Biogeochemistry and Greenhouse Gas Emissions," *Nature Communications* **9**(1348), 1–9. DOI:10.1038/s41467-018- 03352-3.

O'Connell, C. S., et al. 2022. "Utilizing Novel Field and Data Exploration Methods to Explore Hot Moments in High-Frequency Soil Nitrous Oxide Emissions Data: Opportunities and Challenges," *Frontiers in Forests and Global Change* **5**, 674348. DOI:10.3389/ffgc.2022.674348.

Sihi, D., et al. 2021. "Representing Methane Emissions from Wet Tropical Forest Soils Using Microbial Functional Groups Constrained by Soil Diffusivity," *Biogeosciences* **18**(5)1769–86. DOI:10.5194/bg-18-1769-2021.

Silver, W. L., et al. 1999. "Soil Oxygen Availability and Biogeochemistry Along Rainfall and Topographic Gradients in Upland Wet Tropical Forest Soils," *Biogeochemistry* **44**, 301–28. DOI:10.1007/bf00996995.

Venkatakrishnan, S., et al. 2021. "Convolutional Neural Network Based Non-Iterative Reconstruction for Accelerating Neutron Tomography," *Machine Learning Science and Technology* **2**(025031), 1–17.

Xu, X., et al. 2015. "A Microbial Functional Group-Based Module for Simulating Methane Production and Consumption: Application to an Incubated Permafrost Soil," *Journal of Geophysical Research: Biogeosciences* **120**(7), 1315–33. DOI:10.1002/2015jg002935.

Elucidating Environmental Regulation on Microbial-Mediated Soil Methane Emission Using Gene-to-Ecosystem Level Data and Artificial Intelligence

Yang Song
University of Arizona

Focal Areas

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

Science or Technological Challenge

Methane (CH_4) is the second most abundant anthropogenic greenhouse gas after carbon dioxide, accounting for about 20% of global emissions. Although gene-level analysis, ground-level observations, and global-scale Earth system models have separately advanced in interpreting microbial regulation on CH_4 emission and benchmarking global CH_4 emission prediction, the accurate estimation of global CH_4 sources and sinks remains a significant challenge.

Challenge 1. Soil CH_4 emission is driven by specific soil microbes and their released enzymes. However, the spatial distribution of CH_4 emission-associated microbial functions and their interactions with other microbial functions is still unclear. This knowledge gap limits our ability to employ site-level scientific findings to interpret ecosystem-level methane emission uncertainty.

Challenge 2. Soil CH_4 emissions highly fluctuate with spatial heterogeneity and temporal change in multiple environmental factors. Interpreting the nonlinear regulation of multiple environmental factors on soil CH_4 emissions is still difficult, especially considering the potential acclimation and adaptation of soil microbial communities to changing environments. The lack of this knowledge brings significant uncertainty in projecting soil CH_4 emissions under climate change.

Challenge 3. Although there are increased efforts to represent microbial-mediated soil CH_4 emission schemes in the Earth system models (ESMs), the scale difference between the Earth system model and mechanistic understanding of the CH_4 emission process at the gene or lab scale makes it challenging to utilize emerging gene-scale observations to parameterize microbial-mediated soil CH_4 emission schemes in ESMs.

Rationale

Overcoming the above challenges requires performing a regional or global-scale investigation of the distribution of CH_4 emission-associated microbial functions and elucidating how the relative abundances of these microbial functions for CH_4 emission vary with spatiotemporal changes in environmental conditions. Although the emergence of omics technology has brought data to investigate this question, the spatiotemporal representation of these data is still limited. The interpretation of environmental regulation on omics-informed gene function associated with CH_4 emission is highly varied with sampling location. In my previous work, we have harnessed the power of artificial intelligence (AI) and omics data to map soil microbial function involved in soil organic matter decomposition (Flan et al. In review). This study highlights the possibility of synthesizing global-scale omics data to identify microbial functions associated with CH_4 emission and integrating this information with corresponding environmental information to predict the spatiotemporal dynamics of microbial functions for CH_4 emission in response to environmental change. Moreover, to enable the utilization of this gene-scale environmental regulation on microbial CH_4 emission function to advance soil CH_4 flux simulation in ESMs, an effective scaling methodology is required. Microbial-mediated CH_4 emission is an enzyme-catalyzed process and can be calculated using the Michaelis-Menten equation as a function of enzyme abundance, substrate concentration, and kinetics parameters. Therefore, it's possible to utilize the Michaelis-Menten equation to integrate AI prediction of CH_4 emission enzyme functional information with process-based CH_4 flux simulation in ESMs.

Narrative

To overcome current challenges for elucidating the uncertainty from soil CH_4 emission, I propose to develop an integrated research framework that harnesses the power of gene-to-ecosystem scale data, applies AI technology for environment-microbial function prediction and model parameter optimization, and mechanistically advances the representation of soil CH_4 emission in ESMs. In detail, this research frame will need to include: (1) identifying the spatial distribution of microbial enzymes associated with CH_4 emission by integrating omics and environmental information across diverse sites to develop an AI prediction, (2) elucidating environmental regulation on the dynamics of soil enzymes for CH_4 emissions by integrating temporal omics and environmental data to train an AI model for predicting the response of soil enzyme functional composition in response to temporal environmental change, (3) assessing the implication of environmental regulation on CH_4 emissions enzyme for soil CH_4 emission by coupling AI prediction for microbial function for CH_4 emission in response to environmental change with process-based CH_4 flux simulation in the E3SM land model (ELM). The implementation of this research framework will deliver an integrated dataset that pairs microbial CH_4 function information with corresponding environmental information at the global scale. Applying this dataset to AI prediction for the environmental feedback of microbial CH_4 function information will enable us to elucidate climate, edaphic, and vegetation regulation of the composition and abundances of microbial enzyme functions involved in CH_4 emission at the regional or global scale.

Besides this AI application, a surrogate-based AI model will also be employed to optimize new parameters used in omics-informed soil emission simulation in ELM. Applying coupled AI model and ELM prediction will leverage omics data applications in constraining uncertainty in soil CH_4 emission and provide a deep insight into microbial-mediated soil CH_4 emission under changing environments. The success of this study will advance the DOE BER program by overcoming current technical bottlenecks in gene-to-Earth system prediction for the global methane cycle and advance the DOE E3SM model capacity for predicting soil CH_4 emission under more extreme climate conditions.

Reference

Fan, C., et al. “Harness the Power of Machine Learning and Omics to Identify Microbial Functional Composition Across Diverse Environments,” *JGR Biogeoscience.* In review.

AI for Advanced Sensor Data Collection, Automation, and Processing for the Methane Cycle

Maruti K. Mudunuru,[1] Nikolla P. Qafoku,[1] Andre Coleman,[1] Satish Karra,[1] Mariefel Olarte,[1] Sarah Barrows,[1] Tamas Varga,[1] Odeta Qafoku,[1] Glenn Hammond,[1] Behzad Ghanbarian,[2] Mahantesh Halappanavar[1]

[1]Pacific Northwest National Laboratory, [2]Kansas State University

Focal Area

Automated or real-time data capture and processing or federated learning for improvements in measurement coverage.

Science or Technological Challenge

How to use recent advances in AI to obtain automated measurements of methane flux from distributed sensor networks (particularly in soil and agricultural systems) and process the collected data efficiently.

Rationale

Research needs and challenges. One of the main contributors to the methane (CH_4) cycle is the CH_4 gas emitted by microbes in soils that is significantly impacted by human activities (Nazaries et al. 2013). Microorganisms from other sources, such as landfills, livestock, and the exploitation of fossil fuels, also emit CH_4. To better understand methane flux under a wide range of environmental conditions and ecological stressors, various programs (e.g., FLUXNET-CH4, COSORE; Bond-Lamberty et al. 2020; Delwiche et al. 2021) are actively collecting data spatiotemporally that are commonly used in process models (e.g., PFLOTRAN, GCAM; Hammond et al. 2020; Bond-Lamberty et al. 2023) in a coupled modeling-experimental (ModEx) approach. However, there are some challenges associated with this traditional ModEx approach, some of which were recently disclosed within the AI4ESP workshop report highlights (Hickmon et al. 2022). The report highlighted how to use recent advances in AI to overcome some of the traditional ModEx approach challenges. However, many data analysis challenges still need to be answered (Hickmon et al. 2022). Within the context of the methane cycle, we believe there are knowledge gaps that AI would enable us to address by integrating modeling and analysis activities across field- and lab-scale experiments, particularly related to soil and agricultural systems. Those gaps are:

- **Quality of the collected data from sensor networks.** This includes identifying methane flux signatures from sensor data (e.g., microbial activity due to anthropogenic stressors, extreme events), filling in data gaps, and associated data worth analysis.
- **When, how, and where to collect data.** It is not feasible to measure fluxes all the time. Thus, we need a way to decide when, how, and possibly even where to measure methane flux smartly and efficiently.
- **Dealing with big data.** When advanced sensors are used (e.g., multispectral cameras), the amount of data collected is substantial. So, efficiently processing this data at the sensor edge is needed.

Our proposed approach to address these challenges is to develop self-aware and intelligent sensor nodes. This self-awareness is achieved by advancing and tailoring our AI@SensorEdge workflow (e.g., edge-to-cloud intelligence; Mudunuru 2019a, b; Talsma et al. 2023) as shown in Fig. 1, next page.

Narrative

Scientific and technical description. An AI@SensorEdge workflow provides a transformational way to integrate multimodal data through sensor fusion (e.g., combining geophysical, geochemical, and hydrological sensor data sampled at different frequencies). Moreover, efficiently harnessing the connectivity of intelligent sensors through edge and fog computing will result in an advanced understanding of soil and agricultural systems under disturbances and extreme events in near real-time. Development in advanced flux data acquisition systems, sensor network design for soil and farming systems, hardware-related efforts (e.g., AI-enabled accelerators), lightweight AI models (e.g., energy-efficient), and cybersecurity for edge computing will advance the proposed science (Friha et al. 2022). This workflow will recognize:

- **Data quality.** AI-based local data worth analysis will determine if sensor data or signals might contain useful

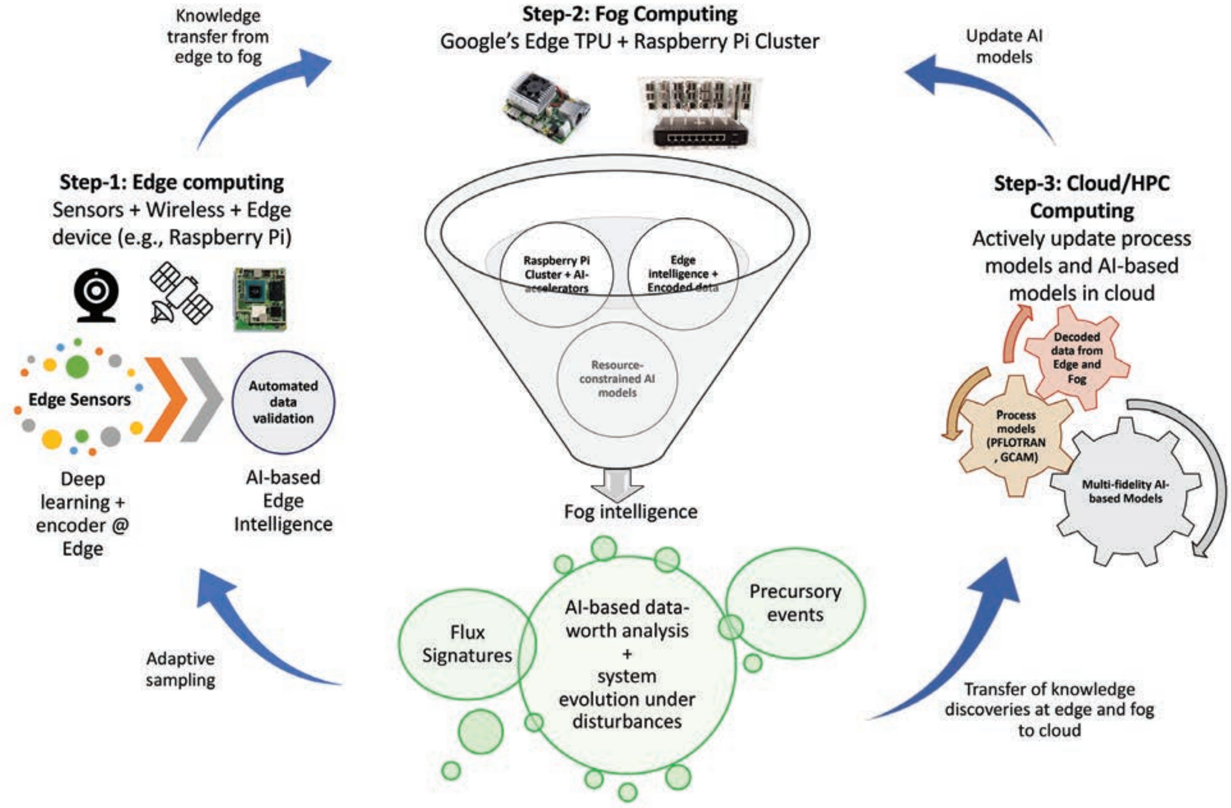

Knowledge transfer from edge to fog
Step-2: Fog Computing
Google's Edge TPU + Raspberry Pi Cluster
Update AI models
Step-1: Edge computing
Sensors + Wireless + Edge device (e.g., Raspberry Pi)
Raspberry Pi Cluster + AI-accelerators
Edge intelligence + Encoded data
Resource-constrained AI models
Step-3: Cloud/HPC Computing
Actively update process models and AI-based models in cloud
Decoded data from Edge and Fog
Process models (PFLOTRAN , GCAM)
Multi-fidelity AI-based Models
Edge Sensors
Automated data validation
Deep learning + encoder @ Edge
AI-based Edge Intelligence
Fog intelligence
Flux Signatures
AI-based data-worth analysis + system evolution under disturbances
Precursory events
Adaptive sampling
Transfer of knowledge discoveries at edge and fog to cloud

Fig. 1. AI@SensorEdge workflow to extract actionable information and discover knowledge from methane flux sensor networks under disturbances. [Credit: Mudunuru et al. 2021]

information to detect flux signatures and underlying patterns. The discovered signatures will be provided to process models (e.g., PFLOTRAN) and then converted to actionable intelligence (e.g., system evolution) at the edge.

- **Data collection.** AI@SensorEdge can accelerate the collection of informative data by creating a digital twin for soil and agricultural systems (e.g., through IoT). We will optimize the location of sensors by exploring the system behavior in digital space.
- **Data volume.** Edge computing-based AI models [e.g., RNNPool (Saha et al. 2020); SmartTensors AI platform (EnviTrace)] can be leveraged to compress data efficiently. This compressed data can be transferred to the cloud and HPC systems through 5G-enabled AI@SensorEdge programming models (Beckman et al. 2020; Argonne National Laboratory, n.d.; University of Chicago, n.d.; Dennis et al. n.d.; TensorFlow Developers 2023).

AI@SensorEdge workflow interfacing with FAIR data sources. The real-time flux measurements collected from sensor networks can be interfaced with existing resources and databases such as:

- **Soil chemistry from Web Soil Survey (NRCS).** To understand the impact of soil chemistry to methane output. www.nrcs.usda.gov/conservation-basics/natural-resource-concerns/soils/soil-geography.
- **Farming data.** For example, types of crops planted in the ground and livestock. quickstats.nass.usda.gov.
- **FLUXNET-CH4.** Methane flux measurements. fluxnet.org/data/fluxnet-ch4-community-product.
- **COSORE.** Soil respiration and greenhouse gas flux data. github.com/bpbond/cosore.
- **Soil Respiration Database.** github.com/bpbond/srdb.

- **Flux measurement sites for carbon, water, and/or energy.** AmeriFlux Network, ameriflux.lbl.gov; Leaf Web, www.leafweb.org; SPRUCE experimental databases, mnspruce.ornl.gov.

Pre-trained AI models can be embedded on these distributed sensor networks through smart computing devices such as Raspberry Pi CM4+. These intelligent edge devices also provide a venue to interface with next generation WiFi and 5G networks. The flux data acquired from these sensor networks and processed using AI algorithms can be made reusable and the findings reproduceable through FAIR data sources such as ESS-DIVE and EMSL-GitHub.

References

Argonne National Laboratory. Waggle: An Edge Computing Platform for Artificial Intelligence and Sensing. https://github.com/waggle-sensor.

Beckman, P., et al. 2020. "5G Enabled Energy Innovation: Advanced Wireless Networks for Science Workshop Report," U.S. Department of Energy Office of Science. DOI:10.2172/1606538.

Bond-Lamberty, B., et al. 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology* **26**(12), 7268–283.

Bond-Lamberty, B., et al. 2023. JGCRI/gcam-core: GCAM 7.0. Zenodo. DOI:10.5281/zenodo.8010145.

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Data* **13**(7), 3607–689.

Dennis, D., et al. *EdgeML: Machine Learning for Resource-Constrained Edge Devices.* Ver. 0.1; microsoft.github.io/EdgeML/.

EnviTrace. *SmartTensors AI Platform*. https://github.com/SmartTensors/.

Friha, O., et al. 2022. "FELIDS: Federated Learning-Based Intrusion Detection System for Agricultural Internet of Things," *Journal of Parallel and Distributed Computing* **165**, 17–31.

Hammond, H., et al. 2020. *PFLOTRAN.* Computer software. U.S. Department of Energy. 2020. Web. DOI:10.11578/dc.20201103.3.

Hickmon, N. L., et al. 2022. "Artificial Intelligence for Earth System Predictability (AI4ESP) 2021 Workshop Report," U.S. Department of Energy Office of Science. Web. DOI:10.2172/1888810.

Mudunuru, M., 2019a. *EDGEip - Intelligent Processing at the Edge to Enhance Efficiency*. DOI:10.13140/RG.2.2.18209.76643.

Mudunuru, M., 2019b. *IoGES: Internet of Things for Geophysical and Environmental Sensing*. DOI:10.13140/RG.2.2.32470.40006.

Nazaries, L., et al. 2013. "Methane, Microbes, and Models: Fundamental Understanding of the Soil Methane Cycle for Future Predictions," *Environmental Microbiology* **15**(9), 2395–2417.

Saha, O., et al. 2020. "RNNPool: Efficient Non-Linear Pooling for RAM Constrained Inference," arXiv:2002.11921[cs.CV].

Talsma, C. J., et al. 2023. "Frost Prediction Using Machine Learning and Deep Neural Network Models for Use on IoT Sensors," *Frontiers in Artificial Intelligence* **5**, 963781. DOI:10.3389/frai.2022.963781.

TensorFlow Developers. 2023. *TensorFlow* v2.12.1.Zenodo. www.tensorflow.org/lite. DOI:10.5281/zenodo.8118033.

University of Chicago. *AoT: Array of Things*. https://arrayofthings.github.io/.

Improving Predictability of Methane Emissions from Terrestrial Ecosystems and Terrestrial-Aquatic Interfaces through Machine Learning Approaches

Debjani Sihi (debjani.sihi@emory.edu)

Emory University

Focal Areas

- This white paper addresses how artificial intelligence (AI) or machine learning (ML) algorithms can improve the predictability of and reduce uncertainties in methane (CH_4) fluxes by integrating complementary data collected at broad spatial and temporal scales.
- We will also address how AI/ML approaches can transfer the knowledge gained from fine (microsite)-scale measurements to field-scale observations and regional- and global-scale budgets.

Science or Technological Challenge

We will focus on processes related to methane (CH_4) production and consumption in terrestrial ecosystems and terrestrial-aquatic interfaces. We will address several data-model integration challenges that directly support BER priorities in enhancing representation of ecosystem processes to improve predictive models. Process-based Earth system models like E3SM lack representations of complex, nonlinear processes related to hot-spots and hot-moments in CH_4 fluxes due to poor understanding of underlying mechanisms related to productions and oxidations of CH_4 (Xu et al. 2016). AI/ML approaches can be used to learn patterns in the data and model errors and use them to inform model structures and equations and correct process-based model errors. Complex, nonlinear processes regulating CH_4 dynamics are difficult to unravel and represent in process-based models. Some correlations may be spurious and not helpful to inform model structure, but AI/ML can help expand human understanding of predictor- response relationships across broad spatial and temporal scales, which, when combined with researcher knowledge, experience, and judgment, can increase our capability to glean insight from complex data. Thus, we can use AI to advance an integrated, robust, and scale-aware predictive understanding of interacting biogeochemical, hydrological, and biophysical processes that enable a new paradigm for improved predictability of CH_4 fluxes.

Rationale

Narrowing uncertainty in regional and global CH_4 budgets is essential for defining necessary policies for climate change mitigation. Significant challenges in reducing uncertainties arise from our incomplete understanding of different underlying processes related to production, consumption, and net fluxes of CH_4 from terrestrial and terrestrial-aquatic interfaces. AI-enabled predictability of methane emissions can close this research gap by cross-scale integration of measurements from multiple sources and disciplines. Insights obtained from laboratory-scale measurements (e.g., dynamics of methanogens or methanotrophs) can inform field-scale observations (e.g., methane fluxes at the biosphere-atmosphere boundary), which could explain patterns in remotely sensed measurements (e.g., atmospheric concentration of CH_4 at regional and global scales).

Narrative

The growing volume of data collected across multiple scales and disciplines offers opportunities to improve AI-enabled predictability of CH_4 fluxes from terrestrial and wetland ecosystems.

Use AI to synthesize automated methane flux measurements and high-frequency sensor data. Technological advances will allow quantification of CH_4 fluxes at sub-daily resolutions (e.g., eddy covariance data and automated chamber data for soil and ecosystem fluxes). Coupling these measurements with *in situ* sensors for soil temperature and moisture can help us identify covarying patterns with seasonal variations and synoptic (i.e., intra-seasonal) oscillations in redox conditions. In 2018, AmeriFlux launched an “Action Theme Year” called Year of Methane (ameriflux.lbl.gov/year-of-methane/year-of-methane/), which brought together the CH_4 flux community to synthesize high-frequency measurements of CH_4 fluxes at the ecosystem scale across FluxNet sites (Knox et al. 2019). Parallel synthesis activities elsewhere resulted in comprehensive datasets of CH_4 fluxes from terrestrial ecosystems (e.g., boreal and Arctic sites, Kuhn et al. 2021) and terrestrial-aquatic interfaces like coastal wetlands (e.g., Coastal Carbon CH_4 working group, serc.si.edu/methane-working-group). Integrating AI/ML algorithms with these continuous measurements at the ecosystem scale, and other community databases like COSORE, can identify contributions of different ecosystem components (e.g., soil, plant) in the net fluxes of CH_4 at the biosphere-atmosphere interface (Megonigal et al. 2008; Bond-Lamberty et al. 2020).

Leverage multi-disciplinary data collected at various spatial and temporal scales to unravel competing mechanisms. Competing processes related to productions and consumptions (or oxidations) of CH_4 can regulate net CH_4 fluxes (Conrad 1989). State-of-the-art techniques, such as isotope pool dilution (von Fischer et al. 2002) and gas push-pull technique (Urmann et al. 2005), are now available to separate net CH_4 fluxes in gross rates of production and consumption in the field. Omics data available from observational networks like MONet (Molecular Observation Network) and NEON (National Ecological Observation Network) can inform spatial variations in microbial functional groups related to CH_4 production (methanogens) and oxidation (methanotrophs) at the continental scale (Xu et al. 2015; Sihi et al. 2021b). Geochemical factors like redox-sensitive elements or alternative electron acceptors (e.g., iron) can further regulate net CH_4 fluxes by influencing the rates of anaerobic oxidation of CH_4 in ecosystems across broad environmental gradients (Teh et al. 2008; Blazewicz et al. 2012; Ettwig et al. 2016; Zheng et al. 2019; Sulman et al. 2022). Synthesis of knowledge obtained from laboratory-scale studies can further quantify the potential effects of microbial, geochemical, and biophysical (redox) processes on observed CH_4 fluxes in the field. Leveraging laboratory, field, and airborne measurements across multiple DOE-funded projects (e.g., NGEE-Tropics, NGEE-Arctic, AmeriFlux, SPRUCE, and COMPASS) can improve our understanding of CH_4 cycle processes in critical ecosystems. AI/ML algorithms can upscale these fine (microsite)-scale measurements of underlying processes to large-scale fluxes by integrating spatial heterogeneity of covarying factors. We expect that implementing explainable ML approaches into the ModEx (model-experimental coupling) framework can improve prediction of hot spots and hot moments (Sihi et al. 2021a) and reduce uncertainty in regional (Zona et al. 2016) and global CH_4 budgets (www.globalcarbonproject.org/methanebudget/).

References

Blazewicz, S. J., et al. 2012. "Anaerobic Oxidation of Methane in Tropical and Boreal Soils: Ecological Significance in Terrestrial Methane Cycling," *Journal of Geophysical Research: Biogeosciences* **117**(G2).

Bond-Lamberty, B., et al. 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology* **26**(12), 7268–283.

Conrad, R. 1989. "Control of Methane Production in Terrestrial Ecosystems," In *Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere.* Wiley, Chichester, U.K. and New York, U.S.A.

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–689.

Ettwig, K. F., et al. 2016. "Archaea Catalyze Iron-Dependent Anaerobic Oxidation of Methane," *PNAS* **113**(45), 12792–796.

Knox, S. H., et al. 2019. "FLUXNET-CH4 Synthesis Activity: Objectives, Observations, and Future Directions," *Bulletin of the American Meteorological Society*, **100**(12), 2607–632.

Kuhn, M. A., et al. 2021. "BAWLD-CH4: A Comprehensive Dataset of Methane Fluxes from Boreal and Arctic Ecosystems," *Earth System Science Data* **13**(11), 5151–189.

Megonigal, J. P., and A. B. Guenther. 2008. "Methane Emissions from Upland Forest Soils and Vegetation," *Tree Physiology* **28**(4), 491–98.

Sihi, D., et al. 2021a. *Improved Understanding of Coupled Water and Carbon Cycle Processes through Machine Learning Approaches.* (No. AI4ESP-1122). Artificial Intelligence for Earth System Predictability (AI4ESP) Collaboration (United States).

Sihi, D., et al. 2021b. "Representing Methane Emissions from Wet Tropical Forest Soils Using Microbial Functional Groups Constrained by Soil Diffusivity," *Biogeosciences* **18**(5), 1769–786.

Sulman, B. N., et al. 2022. "Simulated Hydrological Dynamics and Coupled Iron Redox Cycling Impact Methane Production in an Arctic Soil," *Journal of Geophysical Research: Biogeosciences* **127**(10), e2021JG006662.

Teh, Y. A., et al. 2008. "Suppression of Methanogenesis by Dissimilatory Fe(III)-Reducing Bacteria in Tropical Rain Forest Soils: Implications for Ecosystem Methane Flux," *Global Change Biology* **14**(2), 413–22.

Urmann, K., et al. 2005. "New Field Method: Gas Push-Pull Test for the *In Situ* Quantification of Microbial Activities in the Vadose Zone," *Environmental Science & Technology* **39**(1), 304–10.

von Fischer, J. C., and L. O. Hedin. 2002. "Separating Methane Production and Consumption with a Field-Based Isotope Pool Dilution Technique," *Global Biogeochemical Cycles* **16**(3), 8–1.

Xu, X., et al. 2015. "A Microbial Functional Group-Based Module for Simulating Methane Production and Consumption: Application to an Incubated Permafrost Soil," *Journal of Geophysical Research: Biogeosciences* **120**(7), 1315–333.

Xu, X., et al. 2016. "Reviews and Syntheses: Four decades of Modeling Methane Cycling in Terrestrial Ecosystems," *Biogeosciences* **13**(12), 3735–755.

Zheng, J., et al. 2019. "Modeling Anaerobic Soil Organic Carbon Decomposition in Arctic Polygon Tundra: Insights into Soil Geochemical Influences on Carbon Mineralization," *Biogeosciences* **16**(3), 663–80.

Zona, D., et al. 2016. "Cold Season Emissions Dominate the Arctic Tundra Methane Budget," *PNAS*, **113**(1), 40–45.

Merging Top-Down and Bottom-Up Estimated Wetland CH_4 Emissions Using AI/ML

Sha Feng,[1] Kenneth J. Davis,[2] Anthony Bloom,[3] Z. Jason Hou[1]

[1]Pacific Northwest National Laboratory, [2]Pennsylvania State University, [3]NASA Jet Propulsion Laboratory

Focal Area

When successful, this effort will identify the key mechanisms that govern CH_4 wetland emissions and bridge the gap between top-down and bottom-up estimations.

Science or Technological Challenge

Large discrepancies exist in global CH_4 emission estimations between bottom-up and top-down methods. Since 2012, global CH_4 emissions have been tracking the warmest scenarios assessed by the Intergovernmental Panel on Climate Change (IPCC). Bottom-up methods suggest almost 30% larger global emissions (737 Tg CH_4 / year; range 594–881) than top-down inversion methods. The most important source of uncertainty in the methane budget is attributable to natural emissions, especially those from wetlands and other inland waters (Saunois et al. 2020). AI/ML has successfully brought observational data's insights into model parameterization and calibration. With the accumulation of available flux measurements, there is an opportunity to use AI/ML to bridge the gap between top-down and bottom-up estimated wetland CH_4 emissions.

Rationale

There are numerous factors that contribute to the uncertainty in process-based estimates of CH_4 emissions from wetlands, including model structures, assumptions, parameterization, and selection of forcing data. However, among these sources of uncertainty, the lack of CH_4 flux measurements is a particularly significant factor. In addition, the sensitivity of CH_4 fluxes to environmental controls is not well understood, which also limits explicit representations of many mechanistic processes in models. Top-down methods assimilate atmospheric CH_4 data and have better constraints on emission estimations, but they can only obtain a budget-level estimation of CH_4 emissions without additional information. Our hypothesis is that unknown mechanistic processes hamper the convergence of top-down and bottom-up CH_4 estimations.

Narrative

Large discrepancies exist in global CH_4 emission estimations between bottom-up and top-down methods. With improved partition of wetlands and other inland waters, wetland emissions are about 35 Tg CH_4 /yr lower than previously published budgets (Kirschke et al. 2013; Saunois et al. 2016). However, the overall discrepancy between bottom-up and top-down estimates has been reduced by only 5% compared to Saunois et al. (2016) due to a higher estimate of emissions from inland waters, highlighting the urgent need for an understanding of the mechanisms governing wetland methane emissions.

Bottom-up estimated CH_4 fluxes range from simple empirical models to detailed process-based model simulations, providing prior fluxes for top-down estimations. A process-based model is also the land component of an Earth system model and directly drives climate projections. Previous simulations using process-based models have shown a significant level of uncertainty in estimating wetland CH_4 emissions at regional and global scales. This uncertainty can be attributed to several factors, such as model structures, assumptions, parameterization, and choice of forcing data. Moreover, the impacts of environmental factors on CH_4 fluxes are not entirely clear, which further restricts the explicit representation of various mechanistic processes in models.

Methane wetland emissions are inextricably linked to hydrology. Accordingly, there is considerable intra- and inter-annual variation in emissions in response to variations in precipitation and groundwater. CH_4 model studies face a significant challenge in capturing the complex interactions among climate, soil, and ecosystems. Explicitly representing these interactions in process-based models is difficult without a solid comprehension of the underlying processes.

AI/ML has successfully incorporated insight from observational data into model parameterizations and calibrations. Applying AI/ML to CH_4 wetland models' parameters and variables could further improve CH_4 emission estimations in the domains of recalibration, bias correction, and uncertainty reduction. It is particularly useful in quantifying the responses of nonlinear processes, like CH_4 wetland emissions. With the accumulation of available flux measurements, there is opportunity to use AI/ML to bridge the gap between top-down and bottom-up estimated wetland CH_4 emissions.

Firstly, we use AI/ML methods (e.g., feature selection, dimension reduction, surrogate modeling) to find the key environmental control variables and mechanisms that govern CH_4 fluxes using eddy covariance flux data and the spatially explicit data of climate, hydrology, and soil properties (e.g., soil moisture, temperature, water table level, water storage, etc.).

Secondly, we use AI/ML methods (e.g., smart search, gradient-based, surrogate-assisted, or Bayesian) to optimize parameters associated with key control variables in the CARbon Data MOdel fraMework (CARDAMOM) CH_4 wetland model, and correct CH_4 wetland emission biases. Bloom et al. (2017) developed WetCHARTs, a simple, data-driven, ensemble-based model that produces estimates of CH_4 wetland emissions based on one heterotrophic-respiration model, CARDAMOM, and constrained by observations of precipitation and temperature. CARDAMOM/WetCHARTs will serve as a working surrogate for an Earth system-compliant land model, necessary for us to build the AI/ML capability.

Thirdly, we link atmospheric CH_4 concentrations with CH_4 emissions through an atmospheric transport model and investigate governing processes that affect the temporal and spatial variations of atmospheric CH_4 concentrations. This step could possibly be done by spatiotemporal pattern recognition, AI/ML-based error modeling, and physics-informed AI/ML. The key is to determine the source region of measured atmospheric CH_4 concentrations.

The workflow will be modularized to be easily transferable, generalizable, and efficiently deployable for any terrestrial biogeochemistry model, such as the E3SM land model (ELM).

References

Bloom, A. A., et al. 2017. "A Global Wetland Methane Emissions and Uncertainty Dataset for Atmospheric Chemical Transport Models (WetCHARTs version 1.0)," *Geoscientific Model Development* **10**, 2141–156. DOI:10.5194/gmd-10-2141-2017.

Kirschke, S., et al. 2013. "Three Decades of Global Methane Sources and Sinks," *Nature Geoscience* **6**, 813–23. DOI:10.1038/ngeo1955.

Saunois, M., et al. 2016. "The Global Methane Budget 2000–2012," *Earth System Science Data* **8**, 697–751. DOI:10.5194/essd-8-697-2016.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12**, 1561–1623. DOI:10.5194/essd-12-1561-2020.

Coupling AI-Based Modeling and Molecular Soil Organic Matter at Regional Scale

Satish Karra, Emily Graham, Odeta Qafoku, John R. Bargar, the Environmental Molecular Sciences Laboratory team

Environmental Molecular Sciences Laboratory

Focal Areas

Our white paper is framed around the focal area of the importance of high-potential datasets and how combining multiple datasets leads to scientific insights into the methane cycle. Our approaches in this white paper are also aligned with improving measurement coverage toward reducing uncertainty in mechanistic models.

Science or Technological Challenge

The Environmental Molecular Sciences Laboratory (EMSL) spearheads a 10-year National Molecular Observations Network (MONet; Pacific Northwest National Laboratory) initiative for BER, with the objective to build a national network of environmental sampling and sensing sites along with methods to provide molecular-level and microstructural information on soil, water, resident microbial communities, and biogenic emissions. For instance, in the current phase of the MONet initiative, data types, including metagenomics, respiration, mineral organic matter, hydraulic properties, and geochemistry, are being collected from core samples from a wide range of ecoregions within the United States (EMSL). In coordination and partnership with other observational networks (e.g., ARM, AmeriFlux, NEON), the objective is to make these molecular observations and the data from field-deployed sensors available to the BER community. To make these multi-modal data streams accessible to domain scientists, modelers, and data scientists who study the methane cycle, we aim to build a suite of data and modeling products and avail them to the BER community via the MONet portal. EMSL is strongly positioned to bridge fundamental ModEx gaps by building key products for the BER community. We envision that AI-based methods will be central to these products and are key to accelerating BER community science toward eliciting the mechanics of the methane cycle.

Rationale

Several hydro-biogeochemical natural and anthropogenic processes in the soil, water, and atmosphere, and their complex interactions, contribute to methane fluxes. Characterization of the underlying fundamental molecular-scale and microstructural processes (e.g., geochemistry, omics, etc.) is needed to parameterize and validate the individual process models and their coupling. One of the major contributors to an increase in uncertainty in models is the lack of such data. The MONet initiative at EMSL aims to facilitate the availability of such data to advance model-experiment integration and to enhance the predictive power of multiscale models for carbon and nitrogen fluxes including the methane cycle. Specifically, the key gaps that we will address are:

- Lack of multi-modal molecular and microstructural data with metadata capture that follow FAIR principles for soils across the United States, and the resident microbes and their availability to the BER community.
- Availability of molecular and microstructural data (e.g., analysis, integration, and visualization) and modeling tools (e.g., pore models for transport), along with tools that integrate data and models (e.g., parametrization, sensitivity analysis, uncertainty quantification).
- AI methods can potentially play a major role in these tools and workflows. However, AI methods need data (Gröger 2021) across plot, ecosystem, and regional scales, and the collection of multi-modal molecular and microstructural data is thus needed.

The EMSL MONet soil characterization program, which began user operations in February 2023, provides such molecular data at regional and CONUS scales. MONet is collecting and analyzing soil cores using standardized workflows that can be optimized to provide data critical to AI-informed studies of the methane cycle.

Narrative

Our overall approach is to build a web-based data platform to make MONet observational data, along with AI-based data and modeling tools, available to the BER community.

We briefly detail our vision for the role of AI within data and modeling software products on this platform:

- **AI and graph-based methods for data analysis and visualization.** Classical unsupervised methods, such as principal component analysis (Wang and Zhang 2012), and non-negative matrix factorization (Aittokallio and Schwikowski 2006), have proven to be powerful ways of identifying patterns and dominant features, and in correlating multi-modal datasets. They can be used to identify key signatures in multi-dimensional datasets and reduce dimensionality to visualize data effectively. For instance, our preliminary non-negative matrix factorization analysis on soil biogeochemical and microbial data from EMSL's 1000 Soil Pilot project (a pilot program to MONet), showed clear correlations between dissolved organic matter and environmental stresses, such as flow, pH, and wildfire occurrence. In addition to making unsupervised ML-based tools available, we will build visualization tools based on network theory and graph-based methods for clustering and finding similarities in multi-modal data streams (Tang et al. 2021).
- **AI for multiscale modeling.** To enable the transfer of information (or upscale) from molecular- and microstructural-scale (pore-scale) to the site, regional, and eventual global Earth system models, AI-based methods can play a significant role. For example, MONet will provide users with pore-scale data and models to perform flow and reactive transport simulations, which will then inform averaged parameters, such as reaction rates or permeability, needed in site/regional scale models. AI methods such as deep learning (Ahmmed et al. 2021; Tang et al. 2021) can train on data from such simulations and build surrogate models for upscaling information. These surrogate models will represent the relationships between molecular and microstructural information of interest to the user. Akin to constitutive models or equations of states, AI-based surrogate models can be used in larger-scale simulations.
- **AI for data-model integration.** Recently, AI-based models based on deep learning, including approaches that constrain balance laws (Karra et al. 2021) or mimic balance laws (Haghighat et al. 2021), have become popular. These AI-based models are much faster to run and have been effective for parametrization (Raissi 2018), and quantifying uncertainty (Gasmi and Tchelepi 2022). We will provide users with workflow components that will enable these analyses.

References

Ahmmed, B., et al. 2021. "A Comparative Study of Machine Learning Models for Predicting the State of Reactive Mixing," *Journal of Computational Physics* **432**, 110147.

Aittokallio, T., and B. Schwikowski. 2006. "Graph-Based Methods for Analysing Networks in Cell Biology," *Briefings in Bioinformatics* **7**(3), 243–55. DOI:10.1093/bib/bbl022.

EMSL. "Molecular Observation Network Ecoregions." content-qa.emsl.pnl.gov/sites/default/files/2023-02/MONet_%20Ecoregion%20Table.pdf.

EMSL. "MONet Data Being Collected." https://content-qa.emsl.pnl.gov/sites/default/files/2023- 02/EMSL0419_MonetFlyer.pdf.

Gasmi, C. F., and H. Tchelepi. 2022. "Uncertainty Quantification for Transport in Porous Media Using Parameterized Physics Informed Neural Networks," arXiv:2205.12730[cs.CE].

Gröger, C. 2021. "There is No AI Without Data," *Communications of the ACM* **64**(11), 98–108. DOI:10.1145/3448247.

Haghighat, E., et al. 2021. "A Physics-Informed Deep Learning Framework for Inversion and Surrogate Modeling in Solid Mechanics," *Computer Methods in Applied Mechanics and Engineering* **379**, 113741.

Karra, S., et al. 2021. "AdjointNet: Constraining Machine Learning Models with Physics-Based Code," arXiv:2109.03956[math.NA].

Pacific Northwest National Laboratory. 2021. *Empowering Molecular Discovery Across Scales: EMSL Five-Year Strategic Plan*, PNNL-SA-164144, U.S. Department of Energy Office of Science. https://content-qa.emsl.pnl.gov/sites/default/files/2021-07/EMSLStrategicPlanFY2021_0.pdf.

Raissi, M. 2018. "Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations," *The Journal of Machine Learning Research* **19**(1), 932–55.

Taguchi, Y. H. 2019. *Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach.* Springer International Publishing, New York. DOI:10.1007/978-3-030-22456-1.

Tang, M., et al. 2021. "Deep-Learning-Based Surrogate Flow Modeling and Geological Parameterization for Data Assimilation in 3D Subsurface Flow," *Computer Methods in Applied Mechanics and Engineering* **376**, 113636.

Wang, Y-X., and Y-J. Zhang. 2012. "Nonnegative Matrix Factorization: A Comprehensive Review." *IEEE Transactions On Knowledge and Data Engineering* **25**(6),1336–353. DOI:10.1109/TKDE.2012.51.

Appendix D

References

Adame, F. 2021. "Meaningful Collaborations Can End 'Helicopter Research,'" *Nature.* DOI:10.1038/d41586-021-01795-1.

Argonne National Laboratory. n.d. "Waggle: An Edge Computing Platform for Artificial Intelligence and Sensing." Accessed March 2024. github.com/waggle-sensor.

Barret, M., et al. 2022. "A Combined Microbial and Biogeochemical Dataset from High-Latitude Ecosystems with Respect to Methane Cycle," *Scientific Data* **9**, 674. DOI:10.1038/s41597-022-01759-8.

Bay, S. K., et al. 2021. "Trace Gas Oxidizers are Widespread and Active Members of Soil Microbial Communities," *Nature Microbiology* **6**, 246–256.

Beaulieu, J. J., et al. 2019. "Eutrophication Will Increase Methane Emissions from Lakes and Impoundments During the 21st Century," *Nature Communications* **10**, 1375. DOI:10.1038/s41467-019-09100-5.

Beckman, P., et al. 2020. *5G-Enabled Energy Innovation: Advanced Wireless Networks for Science, Workshop Report.* U.S. Department of Energy Office of Science. DOI:10.2172/1606538.

BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee,* DOE/SC–0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research.

Berg, P., et al. 2022. "Aquatic Eddy Covariance: The Method and Its Contributions to Defining Oxygen and Carbon Fluxes in Marine Environments," *Annual Review of Marine Science* **14**, 431–55. DOI:10.1146/annurev-marine-042121-012329.

Bi, X., et al. 2023. "et*i*Bsu1209: A comprehensive multiscale metabolic model for *Bacillus subtilis*," *Biotechnology and Bioengineering* **120**(6), 1623-39. DOI:10.1002/bit.28355.

Blázquez-García, A., et al. 2020. "A Review on Outlier/Anomaly Detection in Time Series Data," Cornell University. DOI:10.48550/arXiv.2002.04236.

Bond-Lamberty, B., et al. 2020. "COSORE: A Community Database for Continuous Soil Respiration and Other Soil-Atmosphere Greenhouse Gas Flux Data," *Global Change Biology* **26**(12), 7268–83. DOI:10.1111/gcb.15353.

Bond-Lamberty, B., et al. 2021. "A Reporting Format for Field Measurements of Soil Respiration," *Ecological Informatics* **62**, 101280. DOI:10.1016/j.ecoinf.2021.101280.

Borton, B., et al. 2023. "A Functional Microbiome Catalog Crowdsourced from North American Rivers," *BioRxiv.* DOI:10.1101/2023.07.22.550117.

Borton, M., et al. 2022. *GROWdb U.S. River Systems Samples.* U.S. DOE Office of Science, Biological and Environmental Research Program. DOI:10.25982/109073.30/1895615.

Bourgeau-Chavez, L.L., et al. 2021. "Advances in Amazonian Peatland Discrimination with Multi-Temporal PALSAR Refines Estimates of Peatland Distribution, C Stocks, and Deforestation. *Frontiers in Earth Science*, **9**, 676748. DOI:10.3389/feart.2021.676748.

Bousquet, P., et al. 2006. "Contribution of Anthropogenic and Natural Sources to Atmospheric Methane Variability," *Nature* **443**, 439–43. DOI:10.1038/nature05132.

Bridgham, S., et al. 2013. "Methane Emissions from Wetlands: Biogeochemical, Microbial, and Modeling Perspectives from Local to Global Scales," *Global Change Biology* **19**(5), 1325–46. DOI:10.1111/gcb.12131.

Chang, K-Y., et al. 2023. "Observational Constraints Reduce Model Spread but Not Uncertainty in Global Wetland Methane Emission Estimates," *Global Change Biology* **29**(15), 4298–4312. DOI:10.1111/gcb.16755.

Chen, Y., et al. 2022. "Quantifying Regional Methane Emissions in the New Mexico Permian Basin with a Comprehensive Aerial Survey," *Environmental Science and Technology* **56**(7), 4317–23. DOI:10.1021/acs.est.1c06458.

Ciais, P., et al. 2013. "Carbon and Other Biogeochemical Cycles," *Climate Change 2013: The Physical Science Basis.* Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Collier, N., et al. 2018. "The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation," *Journal of Advanced in Modeling Earth Systems* **10**(11), 2731–54. DOI:10.1029/2018MS001354.

Cowan, N., et al. 2020. "Agricultural Soils: A Sink or Source of Methane Across the British Isles?" *European Journal of Soil Science* **72**(4), 1842-1862. DOI:10.1111/ejss.13075.

Crystal-Ornelas, R., et al. 2022a. "Enabling FAIR Data in Earth and Environmental Science with Community-Centric (Meta) Data Reporting Formats," *Scientific Data* **9**, 700. DOI:10.1038/s41597-022-01606-w.

Crystal-Ornelas, R., et al. 2022b. *ESS-DIVE Reporting Format for Location Metadata*. U.S. Department of Energy Office of Science, Biological and Environmental Research Program. DOI:10.15485/1865730.

Dagon, K., et al. 2020. "A Machine Learning Approach to Emulation and Biophysical Parameter Estimation with the Community Land Model, Version 5," *Advances in Statistical Climatology, Meteorology, and Oceanography* **6**(2), 223–44. DOI:10.5194/ascmo-6-223-2020.

Damerow, J., et al. 2021. "Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences," *Data Science Journal* **20**, 11. DOI:10.5334/dsj-2021-011.

da Silva, R. F., et al. 2021. "A Community Roadmap for Scientific Workflows Research and Development," *IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS)*, 81–90. DOI:10.1109/WORKS54523.2021.00016.

Davidson, S. J., et al. 2019. "Wildfire Overrides Hydrological Controls on Boreal Peatland Methane Emissions," *Biogeosciences* **16**(3), 2651–60. DOI:10.5194/bg-16-2651-2019.

Delwiche, K. B., et al. 2021. "FLUXNET-CH4: A Global, Multi-Ecosystem Dataset and Analysis of Methane Seasonality from Freshwater Wetlands," *Earth System Science Data* **13**(7), 3607–89. DOI:10.5194/essd-13-3607-2021.

Dennis, D. K., et al. n.d. "EdgeML: Machine Learning for Resource-Constrained Edge Devices. Ver. 0.1." Accessed March 2024. icrosoft.github.io/EdgeML.

Dentella, V., F. Günther, and E. Leivada. 2023. "Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias," *Proceedings of the National Academy of Sciences* **120**(51), e2309583120. DOI:10.1073/pnas.2309583120.

de Raad, M., et al. 2022. "A Defined Medium for Cultivation and Exometabolite Profiling of Soil Bacteria," *Frontiers in Microbiology* **13**. DOI:10.3389/fmicb.2022.855331.

Dimonaco, N. J., et al. 2022. "No One Tool to Rule Them All: Prokaryotic Gene Prediction Tool Annotations are Highly Dependent on the Organism of Study," *Bioinformatics* **38**(5), 1198–207, DOI:10.1093/bioinformatics/btab827.

Ding, C., et al. 2024. "Mean Annual Precipitation Modulates the Assembly of High-Affinity Methanotroph Communities and Methane Oxidation Activity Across Grasslands," *Agriculture, Ecosystems, and Environment* **360**, 108796.

Dwivedi, D., et al. 2022. "Biogeosciences Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science," *Earth and Space Science* **9**(3), e2021EA002119. DOI:10.1029/2021EA002119.

ElGhawi, R., et al. 2023. "Hybrid Modeling of Evapotranspiration: Inferring Stomatal and Aerodynamic Resistances Using Combined Physics-Based and Machine Learning," *Environmental Research Letters* **18**(3), 034039. DOI:10.1088/1748-9326/acbbe0.

Ellenbogen, J. B., et al. 2023. "Methylotrophy in the Mire: Direct and Indirect Routes for Methane Production in Thawing Permafrost," *mSystems* **9**(1). DOI:10.1128/msystems.00698-23.

Feng, H., et al. 2023. "Global Estimates of Forest Soil Methane Flux Identify a Temperate and Tropical Forest Methane Sink," *Geoderma* **429**(1), 116239. DOI:10.1016/j.geoderma.2022.116239.

Fernandez, R. C., et al. 2023. "How Large Language Models Will Disrupt Data Management," *Proceedings of the VLDB Endowment* **16**(11), 3302–3309. DOI:10.14778/3611479.3611527.

Fluet-Chouinard, E., et al. 2023. "Extensive Global Wetland Loss Over the Past Three Centuries," *Nature* **614**, 281–6. DOI:10.1038/s41586-022-05572-6.

Gatica, G., et al. 2020. "Environmental and Anthropogenic Drivers of Soil Methane Fluxes in Forests: Global Patterns and Among-Biomes Differences," *Global Change Biology* **26**(11), 6604–15. DOI:10.1111/gcb.1533.

Goble, C., et al. 2020. "FAIR Computational Workflows," *Data Intelligence* **2**(1-2), 108–21. DOI:10.1162/dint_a_00033.

Goldman, A. E., et al. 2022. "Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles," *Earth and Space Science* **9**(4), e2021EA002099. DOI:10.1029/2021EA002099.

Guo, J., et al. 2023. "Global Climate Change Increases Terrestrial Soil CH_4 Emissions," *Global Biogeochemical Cycles* **37**(1), e2021GB 007255. DOI:10.1029/2021GB007255.

Hammond, G. E., et al. 2014. "Evaluating the Performance of Parallel Subsurface Simulators: An Illustrative Example with PFLOTRAN," *Water Resources Research* **50**, 208–28.

Hanson, P. J., et al. 2020. "Rapid Net Carbon Loss from a Whole-Ecosystem Warmed Peatland," *AGU Advances* **1**(3), e2020AV000163. DOI:10.1029/2020av000163.

Hargrove, W. W., et al. 2003. "New Analysis Reveals Representativeness of the AmeriFlux Network," *Eos Transactions*, American Geophysical Union. **84**(48), 529–35. DOI:10.1029/2003EO480001.

Hartman, W. H., et al. 2017. "A Genomic Perspective on Stoichiometric Regulation of Soil Carbon Cycling," *ISME Journal* **11**, 2652–65.

He, S., et al. 2015. "Patterns in Wetland Microbial Community Composition and Functional Gene Repertoire Associated with Methane Emissions," *mBio* **6**, e00066-15.

Heděnec, P., et al. 2024. "Global Assessment of Soil Methanotroph Abundances Across Biomes and Climatic Zones: The Role of Climate and Soil Properties," *Applied Soil Ecology* **195**, 105243.

Hoffman, F. M., et al. 2013. "Representativeness-Based Sampling Network Design for the State of Alaska," *Landscape Ecology* **28**(8), 1567–86. DOI:10.1007/s10980-013-9902-0.

Hopple, A. M., et al. 2020. "Massive Peatland Carbon Banks Vulnerable to Rising Temperatures," *Nature Communications* **11**, 2373. DOI:10.1038/s41467-020-16311-8.

Hu, B., et al. 2022. "Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale," *Frontiers in Bioinformatics* **1**, 826370. DOI:10.3389/fbinf.2021.826370.

Huang, Y., et al. 2019. "Realized Ecological Forecast Through an Interactive Ecological Platform for Assimilating Data (Eco-PAD, v1.0) into Models," *Geoscientific Model Development* **12**(3), 1119–37. DOI:10.5194/gmd-12-1119-2019.

IPCC. 2019. "*2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories.*" Ipcc.ch/report/2019-refinement-to-the-2006-ipcc-guidelines-for-national-greenhouse-gas-inventories.

IPCC. 2021. "*Climate Change 2021: The Physical Science Basis. Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.*" Cambridge University Press, Cambridge, United Kingdom, and New York, NY, USA, 2391 pp. DOI: 10.1017/9781009157896.

Irvin, J., et al. 2021. "Gap-Filling Eddy Covariance Methane Fluxes: Comparison of Machine Learning Model Predictions and Uncertainties at FLUXNET-CH4 Wetlands," *Agricultural and Forest Meteorology* **308–09**, 108528. DOI:10.1016/j.agrformet.2021.108528.

Jansen, J., et al. 2022. "Global Increase in Methane Production Under Future Warming of Lake Bottom Waters," *Global Change Biology* **28**(18), 5427-5440. DOI:10.1111/gcb.16298.

Jing, H., et al. 2020. "Anaerobic Methane Oxidation Coupled to Denitrification Is an Important Potential Methane Sink in Deep-Sea Cold Seeps," *Science of the Total Environment* **748**, 142459. DOI:10.1016/j.scitotenv.2020.142459.

Johnson, M. S., et al. 2022. "Methane Emission from Global Lakes: New Spatiotemporal Data and Observation-Driven Modeling of Methane Dynamics Indicates Lower Emissions," *JGR Biogeosciences* **127**(7), e2022JG006793.

Jumper, J., et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* **596**, 583–589. DOI: 10.1038/s41586-021-03819-2.

Jung, M., et al. 2011. "Global Patterns of Land-Atmosphere Fluxes of Carbon Dioxide, Latent Heat, and Sensible Heat Derived from Eddy Covariance, Satellite, and Meteorological Observations," *Journal of Geophysical* Research **116**, G00J07. DOI:10.1029/2010JG001566.

Jung, M., et al. 2019. "The FLUXCOM Ensemble of Global Land-Atmosphere Energy Fluxes," *Scientific Data* **6**, 74. DOI:10.1038/s41597-019-0076-8.

Kavvas, E. S., et al. 2020. "A Biochemically Interpretable Machine Learning Classifier for Microbial GWAS," *Nature Communications* **11**, 2580. DOI:10.1038/s41467-020-16310-9.

Keller, M., et al. 2008. "A Continental Strategy for the National Ecological Observatory Network," *Frontiers in Ecology and Environment* **6**(5), 282–84. DOI:10.1890/1540-9295(2008)6[282: ACSFTN]2.0.CO;2.

Keremedjiev, M., et al. 2022. "Carbon Mapper Phase 1: Two Upcoming VNIR-SWIR Hyperspectral Imaging Satellites," *Proceedings SPIE 12094, Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXVIII*, 1209409. DOI:10.1117/12.2632547.

Ketzer, M., et al. 2020. "Gas Hydrate Dissociation Linked to Contemporary Ocean Warming in the Southern Hemisphere," *Nature Communications* **11**, 3788. DOI:10.1038/s41467-020-17289-z.

Khan, M. A. W., et al. 2023. "Amazonian Soil Metagenomes Indicate Different Physiological Strategies of Microbial Communities in Response to Land Use Change," *Applied Environmental Microbiology*. Under revision.

Kim, K., E. J. Daly, and G. Hernandez-Ramirez. 2021. "Perennial Grain Cropping Enhances the Soil Methane Sink in Temperate Agroecosystems," *Geoderma* **388**, 114931. DOI:10.1016/j.geoderma.2021.11493.

Kirschke, S., et al. 2013. "Three Decades of Global Methane Sources and Sinks," *Nature Geoscience* **6**, 813–23. DOI:10.1038/ngeo1955.

Knox, S. H., et al. 2019. "FLUXNET-CH4 Synthesis Activity: Objectives, Observations, and Future Directions," *Bulletin of American Meteorological Society* **100**(12), 2607–32. DOI:10.1175/BAMS-D-18-0268.1.

Knox, S. H., et al. 2021. "Identifying Dominant Environmental Predictors of Freshwater Wetland Methane Fluxes Across Diurnal to Seasonal Time Scales," *Global Change Biology* **27**(15), 3582–604.

Koffi, E. N., et al. 2020. "An Observation-Constrained Assessment of the Climate Sensitivity and Future Trajectories of Wetland Methane Emissions," *Science Advances* 6(15), eaay4444. DOI:10.1126/sciadv.aay4444.

Kuhn, M. A., et al. 2021. "BAWLD-CH_4: A Comprehensive Dataset of Methane Fluxes from Boreal and Arctic Ecosystems," *Earth System Science Data* **13**, 5151–89. DOI:10.5194/essd-13-5151-2021.

Kumar, J., et al. 2016. "Understanding the Representativeness of FLUXNET for Upscaling Carbon Flux from Eddy Covariance Measurements," *Earth System Science Data*, 1–25. DOI:10.5194/essd-2016-36.

Lacroix, E. M., et al. 2023. “Consider the Anoxic Microsite: Acknowledging and Appreciating Spatiotemporal Redox Heterogeneity in Soils and Sediments,” *ACS Earth and Space Chemistry* **7**(9), 1592–1609.

Lai, D. Y. F. 2009. “Methane Dynamics in Northern Peatlands: A Review,” *Pedosphere* **19**(4), 409–21. DOI:10.1016/S1002-0160(09)00003-4.

Lan, X., et al. 2021. “What Do We Know About the Global Methane Budget? Results from Four Decades of Atmospheric CH_4 Observations and the Way Forward,” *Philosophical Transactions of the Royal Society A* **379**, 20200440. DOI: 0.1098/rsta.2020.0440.

Larsen, P. E., et al. 2011. “Predicted Relative Metabolomic Turnover (PRMT): Determining Metabolic Turnover from a Coastal Marine Metagenomic Dataset,” *Microbial Informatics and Experimentations* **1**, 4. DOI:10.1186/2042-5783-1-4.

Lewis, P., et al. 2021. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv:2005.11401v4 [cs.CL].

Liu, F., et al. 2023. “AI4CH4: Building Mechanistic Understanding of Environmental Microbiomes.” Argonne National Laboratory. Accessed March 2024. youtube.com/watch?v=1g-pnCKRusw.

Lu, D., and D. Ricciuto. 2019. “Efficient Surrogate Modeling Methods for Large-Scale Earth System Models Based on Machine-Learning Techniques,” *Geoscientific Model Development* **12**(5), 1791–807. DOI:10.5194/gmd-12-1791-2019.

Lu, D., et al. 2022. “An Interpretable Machine Learning Model for Advancing Terrestrial Ecosystem Predictions,” *The International Conference on Learning Representations.*

Malakhova, V., and E. Golubeva. 2022. “Model Study of the Effects of Climate Change on the Methane Emissions on the Arctic Shelves,” *Atmosphere* **13**(2), 274. DOI:10.3390/atmos13020274.

Malhotra, A., and N. T. Roulet. 2015. “Environmental Correlates of Peatland Carbon Fluxes in a Thawing Landscape: Do Transitional Thaw Stages Matter?” *Biogeosciences* **12**(10), 3119–30.

Malhotra, A., et al. 2020. “Peatland Warming Strongly Increases Fine-Root Growth,” *Proceedings of the National Academy of Sciences USA* **117**(30), 17627–34.

Mallick, H., et al. 2019. “Predictive Metabolomic Profiling of Microbial Communities Using Amplicon or Metagenomic Sequences,” *Nature Communications* **10**, 3136.

McNicol, G., et al. 2023. “Upscaling Wetland Methane Emissions from the FLUXNET-CH4 Eddy Covariance Network ($UpCH_4$ v1.0): Model Development, Network Assessment, and Budget Comparison,” *AGU Advances* **4**(5), e2023AV000956. DOI:10.1029/2023AV000956.

Megonigal, J., et al. 2004. “Anaerobic Metabolism: Linkages to Trace Gases and Aerobic Processes,” *Biogeochemistry*, 317–424.

Metcalfe, K. S., et al. 2021. “Experimentally Validated Correlation Analysis Reveals New Anaerobic Methane Oxidation Partnerships with Consortium-Level Heterogeneity in Diazotrophy,” *ISME Journal* **15**, 377–396. DOI:10.1038/s41396-020-00757-1.

Mishra, U., et al. 2020. “Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region,” *Frontiers in Big Data*, **3**(40). DOI:10.3389/fdata.2020.528441.

Mishra, U., et al. 2021. “Spatial Heterogeneity and Environmental Predictors of Permafrost Region Soil Organic Carbon Stocks,” *Science Advances* **7**(9). DOI:10.1126/sciadv.aaz5236.

Mital, U., et al. 2020. “Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests,” *Frontiers in Water* **2**. DOI:10.3389/frwa.2020.00020.

Moon, M., et al. 2022. “A High Spatial Resolution Land Surface Phenology Dataset for AmeriFlux and NEON Sites,” *Scientific Data* **9**, 448. DOI: 10.1038/s41597-022-01570-5.

Morin, T. H., et al. 2017. “Combining Eddy-Covariance and Chamber Measurements to Determine the Methane Budget from a Small, Heterogeneous Urban Floodplain Wetland Park,” *Agricultural and Forest Meteorology* **237–238**, 160–170. DOI:10.1016/j.agrformet.2017.01.022.

Moxon, S., et al. 2021. “The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics,” *CEUR Workshop Proceedings* **3073**, 148–151. Accessed March 2024. pure.johnshopkins.edu/en/publications/the-linked-data-modeling-language-linkml-a-general-purpose-data-m.

Mudunuru, M. K., et al. 2021. *EdgeAI: How to Use AI to Collect Reliable and Relevant Watershed Data, AI4ESP-1095*. U.S. Department of Energy Office of Science, Biological and Environmental Research Program. DOI:10.2172/1769700.

Müller, J., et al. 2015. “CH_4 Parameter Estimation in CLM4.5bgc Using Surrogate Global Optimization,” *Geoscientific Model Development* **8**(10), 3285–310. DOI:10.5194/gmd-8-3285-2015.

Murguia-Flores, F., et al. 2021. “Global Uptake of Atmospheric Methane by Soil From 1900 to 2100,” *Global Biogeochemical Cycle* **35**(7), e2020GB006774. DOI:10.1029/2020GB006774.

Narrowe, A. B., et al. 2019. “Uncovering the Diversity and Activity of Methylotrophic Methanogens in Freshwater Wetland Soils,” *mSystems* **4**(6), 10.1128/msystems.00320-19.

National Oceanic and Atmospheric Administration (NOAA). 2022. “Increase in Atmospheric Methane Set Another Record During 2021.” Accessed March 2024. Noaa.gov/news-release/increase-in-atmospheric-methane-set-another-record-during-2021.

Nguyen, T., et al. 2023. “ClimaX: A Foundation Model for Weather and Climate.” arXiv:2301.10343v5 [cs.LG].

Ni, X., and P. Groffman. 2018. "Declines in Methane Uptake in Forest Soils," *Proceedings of the National Academy of Sciences USA* **115**(34), 8587–90. DOI:10.1073/pnas.1807377115.

Nozhevnikova, A. N., et al. 2020. "Syntrophy and Interspecies Electron Transfer in Methanogenic Microbial Communities," *Microbiology* **89**, 129–147. DOI:10.1134/S0026261720020101.

Noyce, G. L., et al. 2019. "Asynchronous Nitrogen Supply and Demand Produce Nonlinear Plant Allocation Responses to Warming and Elevated CO_2," *Proceedings of the National Academy of Sciences USA* **116**(43), 21623–28. DOI:10.1073/pnas.1904990116.

Noyce, G. L., et al. 2023. "Oxygen Priming Induced by Elevated CO_2 Reduces Carbon Accumulation and Methane Emissions in Coastal Wetlands," *Nature Geoscience* **16**(1), 63–68. DOI:10.1038/s41561-022-01070-6.

Oh, Y., et al. 2016. "A Scalable Model for Methane Consumption in Arctic Mineral Soils," *Geophysical Research Letters* **43**(10), 5143–5150. DOI: 10.1002/2016GL069049.

Oh, Y., et al. 2020. "Reduced Net Methane Emissions Due to Microbial Methane Oxidation in a Warmer Arctic," *Nature Climate Change* **10**, 317–321. DOI: 10.1038/s41558-020-0734-z.

Orphan, V. J., et al. 2022. *DOE New Players Carbon Cycle (2016-2021)*. U.S. DOE Office of Science, Biological and Environmental Research Program, Biological Systems Science Division. osti.gov/biblio/1872276.

Øyås, O., et al. 2024. "Predicting Microbial Genome-Scale Metabolic Networks Directly from 16S rRNA Gene Sequences," *bioRxiv*. DOI:10.1101/2024.01.26.576649.

Pallandt, M. M. T. A., et al. 2022. "Representativeness Assessment of the Pan-Arctic Eddy Covariance Site Network and Optimized Future Enhancements," *Biogeosciences* **19**(3), 559–583. DOI:10.5194/bg-19-559-2022.

Park, J., et al. 2023. "Long-Term Missing Value Imputation for Time Series Data Using Deep Neural Networks," *Neural Computing and Applications* **35**, 9071–91. DOI:10.1007/s00521-022-08165-6.

Pastorello, G., et al. 2020. "The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data," *Scientific Data* **7**, 225. DOI:10.1038/s41597-020-0534-3.

Pausas, J. G., and J. E. Keeley. 2021. "Wildfires and Global Change," *Frontiers in Ecology and the Environment* **21**(8), 387–95. DOI:10.1002/fee.2359.

Peng, S., et al. 2022. "Wetland Emission and Atmospheric Sink Changes Explain Methane Growth in 2020," *Nature* **612**, 477–82. DOI:10.1038/s41586-022-05447-w.

Pham-Duc, B., et al. 2017. "Comparisons of Global Terrestrial Surface Water Datasets Over 15 Years." *Journal of Hydrometeorology* **18**, 993–1007. DOI:10.1175/JHM-D-16-0206.1.

Pi, X., et al. 2022. "Mapping Global Lake Dynamics Reveals the Emerging Roles of Small Lakes," *Nature Communications* **13**(1), 5777. DOI:10.1038/s41467-022-33239-3.

Quebbeman, A. W., et al. 2022. "A Severe Hurricane Increases Carbon Dioxide and Methane Fluxes and Triples Nitrous Oxide Emissions in a Tropical Forest," *Ecosystems* **25**, 1754–66. DOI:10.1007/s10021-022-00794-1.

Ramachandran, R., et al. 2022. "Language Model for Earth Science: Exploring Potential Downstream Applications as Well as Current Challenges," *IEEE Xplore*. DOI:10.1109/IGARSS46834.2022.9883682.

Randerson, J. T., et al. 2009. "Systematic Assessment of Terrestrial Biogeochemistry in Coupled Climate-Carbon Models," *Global Change Biology* **15**(10), 2462–84. DOI:10.1111/j.1365-2486.2009.01912.x.

Reiman, D., B. T. Layden, and Y. Dai. 2021. "MiMeNet: Exploring Microbiome-Metabolome Relationships Using Neural Networks," *PLOS Computational Biology* **17**(5): e1009021. DOI:10.1371/journal.pcbi.1009021.

Reinhard, C. T., et al. 2020. "Oceanic and Atmospheric Methane Cycling in the cGENIE Earth System Model, release Ver. 0.9.14," *Geoscientific Model Development* **13**(11), 5687–5706. DOI: 0.5194/gmd-13-5687-2020.

Reisinger, A., et al. 2021. "How Necessary and Feasible Are Reductions of Methane Emissions from Livestock to Support Stringent Temperature Goals?" *Philosophical Transactions of the Royal Society A* **379**(2210), 20200452. DOI: 10.1098/rsta.2020.0452.

Ricciuto, D., et al. 2021. "An Integrative Model for Soil Biogeochemistry and Methane Processes: I. Model Structure and Sensitivity Analysis," *Journal of Geophysical Research-Biogeosciences* **126**(8), e2019JG005468. DOI:10.1029/2019JG005468.

Riley, W. J., et al. 2011. "Barriers to Predicting Changes in Global Terrestrial Methane Fluxes: Analyses Using CLM4Me, a Methane Biogeochemistry Model Integrated in CESM," *Biogeosciences* **8**(7), 1925–53. DOI:10.5194/bg-8-1925-2011.

Rößger, N., et al. 2022. "Seasonal Increase of Methane Emissions Linked to Warming in Siberian Tundra," *Nature Climate Change* **12**, 1031–1036. DOI:10.1038/s41558-022-01512-4.

Rodriguez, L. K., et al. 2023. "LAGOS-U.S. RESERVOIR: A Database Classifying Conterminous U.S. Lakes 4 ha and Larger as Natural Lakes or Reservoir Lakes." Limnology and Oceanography Letters, **8**(2), 267–285.

Rogelj, J. and R. D. Lamboll. 2024. "Substantial Reductions in Non-CO_2 Greenhouse Gas Emissions Reductions Implied by IPCC Estimates of the Remaining Carbon Budget," *Communications Earth and Environment* **5**, 35. DOI: 10.1038/s43247-023-01168-8.

Rosentreter, J. A., et al. 2021. "Half of Global Methane Emissions Come from Highly Variable Aquatic Ecosystem Sources," *Nature Geoscience* **14**, 225–30. DOI:10.1038/s41561-021-00715-2.

Runkle, B. R. K., et al. 2019. "Methane Emission Reductions from the Alternate Wetting and Drying of Rice Fields Detected Using the Eddy Covariance Method," *Environmental Science & Technology* **53**(2), 671–81. DOI:10.1021/acs.est.8b05535.

Ryu, S., et al. 2020. "Denoising Autoencoder-Based Missing Value Imputation for Smart Meters," *The Institute of Electrical and Electronics Engineers Access* **8**, 40656–66. DOI:10.1109/ACCESS.2020.2976500.

Sargsyan, K., et al. 2014. "Dimensionality Reduction for Complex Models Via Bayesian Compressive Sensing," *International Journal for Uncertainty Quantification* **4**(1), 63–93. DOI:10.1615/Int.J.UncertaintyQuantification.201300682.

Saunois, M., et al. 2020. "The Global Methane Budget 2000–2017," *Earth System Science Data* **12**(3), 1561–623. DOI:10.5194/essd-12-1561-2020.

Schimel, D., et al. 2007. "NEON: A Hierarchically Designed National Ecological Network," *Frontiers in Ecology and Environment* **5**(2), 59. DOI:10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2.

Schuur, E., et al. 2022. "Permafrost and Climate Change: Carbon Cycle Feedbacks from the Warming Arctic," *Annual Review of Environment and Resources* **47**(1), 343–71. DOI:10.1146/annurev-environ-012220-011847.

Sihi, D., et al. 2021. "Representing Methane Emissions from Wet Tropical Forest Soils Using Microbial Functional Groups Constrained by Soil Diffusivity," *Biogeosciences* **18**, 1769–86. DOI:10.5194/bg-18-1769-202.

Simmonds, M. B., et al. 2022. "Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis," *Data Science Journal* **21**, 3. DOI:10.5334/dsj-2022-003.

Skennerton, C. T., et al. 2017. "Methane-Fueled Syntrophy through Extracellular Electron Transfer: Uncovering the Genomic Traits Conserved within Diverse Bacterial Partners of Anaerobic Methanotrophic Archaea," *mBio* **8**(4), 10.1128/mbio.00530-17.

Smith, G. J., and K. C. Wrighton. 2019. "Metagenomic Approaches Unearth Methanotroph Phylogenetic and Metabolic Diversity," *Current Issues in Molecular Biology* **33**, 57–84. DOI:10.21775/cimb.033.057.

Song, C., et al. 2020. "A Microbial Functional Group-Based CH_4 Model Integrated into a Terrestrial Ecosystem Model: Model Structure, Site-Level Evaluation, and Sensitivity Analysis," *Journal of Advances in Modeling Earth Systems* **12**(4), e2019MS001867. DOI:10.1029/2019MS001867.

Sturtevant, C., et al. 2016. "Identifying Scale-Emergent, Nonlinear, Asynchronous Processes of Wetland Methane Exchange," *JGR Biogeosciences* **121**(1), 188–204.

Sutton-Grier, A. E., and J. P. Megonigal. 2011. "Plant Species Traits Regulate Methane Production in Freshwater Wetland Soils," *Soil Biology and Biochemistry* **43**(2), 413–20.

Takahashi, H., et al. 2013. "Aerenchyma Formation in Plants," *Low-Oxygen Stress in Plants* **21**, 247–65. DOI:10.1007/978-3-7091-1254-0_13.

Tang, L. 2023. "Large Models for Genomics," *Nature Methods* **20**, 1868. DOI:10.1038/s41592-023-02105-5.

TensorFlow Developers. 2023. "TensorFlow v2.12.1.Zenodo." DOI:10.5281/zenodo.8118033. tensorflow.org/lite.

Todd-Brown, K. E. O., et al. 2022. "Reviews and Syntheses: The Promise of Big Diverse Soil Data, Moving Current Practices Towards Future Potential," *Biogeosciences* **19**(14), 3505–22.

Toro, S., et al. 2023. "Dynamic Retrieval Augmented Generation of Ontologies Using Artificial Intelligence (DRAGON-AI)," arXiv:2312.10904v1 [cs.AI].

Turner, A., et al. 2019. "Interpreting Contemporary Trends in Atmospheric Methane," *Proceedings of the National Academy of Sciences* **116**(8), 2805–13. DOI:10.1073/pnas.1814297116.

Ueyama, M., et al. 2022. "Partitioning Methane Flux by the Eddy Covariance Method in a Cool Temperate Bog Based on a Bayesian Framework," *Agricultural and Forest Meteorology* **316**, 108852. DOI:10.1016/j.agrformet.2022.108852.

Ueyama, M., et al. 2023. "Modeled Production, Oxidation, and Transport Processes of Wetland Methane Emissions in Temperate, Boreal, and Arctic Regions," *Global Change Biology* **29**(8), 2313–2334. DOI:10.1111/gcb.16594.

United Nations Environment Programme (UNEP). 2023. *An Eye on Methane: The Road to Radical Transparency. International Methane Emissions Observatory 2023 Report.* ISBN: 978-92-807-4102-5.

University of Chicago. n.d. "AoT: Array of Things Technology." Accessed March 2024. Arrayofthings.github.io.

U.S. DOE. 2018. *Earth and Environmental Systems Sciences Division Strategic Plan 2018–2023*, DOE/SC–0192. U.S. Department of Energy Office of Science.

U.S. DOE. 2020. *AI for Science*. U.S. Department of Energy Office of Science. Anl.gov/cels/reference/ai-for-science-report-2020.

U.S. DOE. 2022. *Artificial Intelligence for Earth System Predictability (AI4ESP)*. U.S. Department of Energy Office of Science. DOI:10.2172/1888810.

U.S. DOE. 2023. *Artificial Intelligence and Machine Learning for Bioenergy Research: Opportunities and Challenges*, DOE/SC-0211. U.S. Department of Energy Office of Science and Office of Energy Efficiency and Renewable Energy. DOI:10.2172/1968870.

van Bodegom, P., et al. 2001. "Methane Oxidation and the Competition for Oxygen in the Rice Rhizosphere," *Applied and Environmental Microbiology* **67**(8), 3586–97. DOI:10.1128/AEM.67.8.3586-3597.2001.

Varadharajan, C., and H. F. Hemond. 2012. "Time-Series Analysis of High-Resolution Ebullition Fluxes from a Stratified, Freshwater Lake," *Journal of Geophysical Research–Biogeosciences* **117**(G2). DOI:10.1029/2011JG001866.

Varadharajan, C., et al. 2021. "BASIN-3D: A Brokering Framework to Integrate Diverse Environmental Data," *Computers & Geosciences* **159**, 105024. DOI:10.1016/j.cageo.2021.105024.

Varadi, M., et al. 2022. "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models," *Nucleic Acids Research* **50**(D1), D439-D444. DOI:10.1093/nar/gkab1061.

Vigderovich, H., et al. 2023. "Aerobic Methanotrophy Increases the Net Iron Reduction in Methanogenic Lake Sediments," *Frontiers in Microbiology* **14,** 1206414. DOI:10.3389/fmicb.2023.1206414.

Volkova, L., et al. 2014. "Fuel Reduction Burning Mitigates Wildfire Effects on Forest Carbon and Greenhouse Gas Emission," *International Journal of Wildland Fire* **23**(6), 771–80. DOI:10.1071/ WF14009.

Wany, A., and K. J. Gupta. 2018. "Reactive Oxygen Species, Nitric Oxide Production and Antioxidant Gene Expression During Development of Aerenchyma Formation in Wheat," *Plant Signaling and Behavior* **13**(2), e1428515. DOI:10.1080/15592324.2018.1428515.

Warren, A. S., et al. 2010. "Missing Genes in the Annotation of ProKaryotic Genomes," *BMC Bioinformatics* **11**, 131. DOI:10.1186/1471-2105-11-131.

Whiting, G. J., and J. P. Chanton. 1993. "Primary Production Control of Methane Emission from Wetlands," *Nature* **364**, 794–95. DOI:10.1038/364794a0.

Wilkinson, M. D., et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* **3**(60018). DOI:10.1038/sdata.2016.18.

Wilkinson, S. L., et al. 2023. "Wildfire and Degradation Accelerate Northern Peatland Carbon Release," *Nature Climate Change* **13**, 456–61. DOI:1038/s41558-023-01657-w.

Wu, X., et al. 2020. "Culturing of 'Unculturable' Subsurface Microbes: Natural Organic Carbon Source Fuels the Growth of Diverse and Distinct Bacteria from Groundwater," *Frontiers in Microbiology* **11**. DOI: 10.3389/fmicb.2020.610001.

Xu, X., et al. 2015. "A Microbial Functional Group Based Module for Simulating Methane Production and Consumption: Application to an Incubation Permafrost Soil," *Journal of Geophysical Research–Biogeosciences* **120**, 1315–33.

Xu, X., et al. 2016. "Review and Synthesis: Four Decades of Modeling Methane Cycling Within Terrestrial Ecosystems," *Biogeosciences* **13,** 3735–55. DOI:10.5194/bg-13-3735-2016.

Yilmaz, P., et al. 2011. "Minimum Information About a Marker Gene Sequence (MIMARKS) and Minimum Information About Any (X) Sequence (MIxS) Specifications," *Nature Biotechnology* **29**, 415–20. DOI:10.1038/nbt.1823.

Yoo, Y. H., et al. 2015. "Genome-Wide Identification and Analysis of Genes Associated with Lysigenous Aerenchyma Formation in Rice Roots," *Journal of Plant Biology* **58**, 117–27. DOI:10.1007/s12374-014-0486-2.

Yuan, F., et al. 2021. "An Integrative Model for Soil Biogeochemistry and Methane Processes II, Warming and Elevated CO_2 Impacts on Peatland CH_4 Emission," *Journal of Geophysical Research–Biogeosciences* **126**(8), e2020JG005963. DOI:10.1029/2020JG005963.

Yuan, F., et al. 2023. "Evaluation and Improvement of the E3SM Land Model for Simulating Energy and Carbon Fluxes in an Amazonian Peatland," *Agricultural and Forest Meteorology* **332**, 109364. DOI:10.1016/j.agrformet.2023.109364.

Yuan, K., et al. 2022. "Causality Guided Machine Learning Model on Wetland CH_4 Emissions Across Global Wetlands," *Agricultural and Forest Meteorology* **324**, 109115. DOI:10.1016/j.agrformet.2022.109115.

Zhang, Z., et al. 2021. "Development of the Global Dataset of Wetland Area and Dynamics for Methane Modeling (WAD2M)," Earth System Science Data **13**(5), 2001–2023. DOI:10.5194/essd-13-2001-2021.

Zhang, Z., et al. 2023. "Recent Intensification of Wetland Methane Feedback," *Nature Climate Change* **13**, 430–33. DOI:10.1038/s41558-023-01629-0.

Zhao, J. F., et al. 2019. "Tropical Forest Soils Serve as Substantial and Persistent Methane Sinks," *Scientific Reports* **9**, 16799. DOI:10.1038/s41598-019-51515-z.

Zhu, Q., et al. 2022. "Building a Machine Learning Surrogate Model for Wildfire Activities within a Global Earth System Model," *Geoscientific Model Development* **15**(5), 1899–911. DOI:10.5194/gmd-15-1899-2022.

Zhuang, Q., et al. 2013. "Response of Global Soil Consumption of Atmospheric Methane to Changes in Atmospheric Climate and Nitrogen Deposition," *Global Biogeochemical Cycles* **27**(3), 650–63. DOI:10.1002/gbc.20057.

Zhuang, Q., et al. 2023. "Current and Future Global Lake Methane Emissions: A Process-Based Modeling Analysis," *JGR Biogeosciences* **128**(3), e2022JG007137. DOI:10.1029/2022JG007137.

Appendix E

Image Credits

Image captions and credits for Fig. 4.1, p. 26.

Satellite orbiting over Earth. [Courtesy Adobe Stock]

Niwot Ridge AmeriFlux tower near Nederland, Colo. [Courtesy AmeriFlux]

The Atomospheric Radiation Measurement (ARM) user facility employs this G-1 aircraft to measure a range of aerosol and cloud properties, as well as collect gas-phase measurements. [Courtesy U.S. Department of Energy Atmospheric Radiation Measurement (ARM) user facility.]

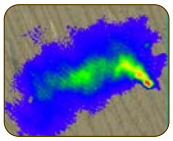
Aerial drone. [Courtesy Adobe Stock]

Eddy covariance system deployed at the Atmospheric Radiation Measurement (ARM) Central Facility and operated under the ARM Carbon Project umbrella at Lawrence Berkeley National Laboratory. [Courtesy Lawrence Berkeley National Laboratory]

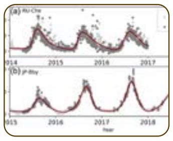
A researcher in a canoe uses a floating chamber and a gas analyzer to measure methane emission from Old Woman Creek wetland in Ohio. [Courtesy The Ohio State University]

A researcher extracts DNA from a field-collected wetland sediment sample. [Courtesy Colorado State University]

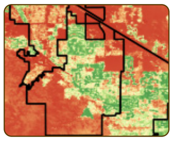
A scientist sequences genetic material at the DOE Joint Genome Institute (JGI). [Courtesy JGI]



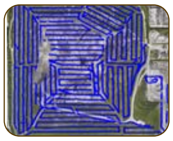
Methane leak detection and measurement via satellite. [Courtesy GHGSat]



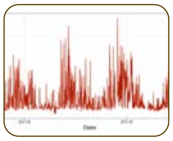
A graph indicates methane emission increasing over time as measured by field chambers. [Reprinted with permission from Elsevier from Villa, J. A., et al. 2021. "Ebullition Dominates Methane Fluxes from the Water Surface Across Different Ecohydrological patches in a Temperate Freshwater Marsh at the End of the Growing Season," *Science of the Total Environment* **767**, 14498. DOI:10.1016/j.scitotenv.2020.144498.]



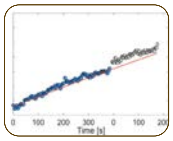
Evaporative Stress Index at Coachella Valley, Calif., derived from ECOSTRESS. [Courtesy NASA]



Drone-based surface emission monitoring data is utilized to map methane emissions. Reprinted under a Creative Commons Attribution (CC BY) license from Abichou, T., et al. 2023. "Using Ground- and Drone-Based Surface Emission Monitoring (SEM) Data to Locate and Infer Landfill Methane Emissions," *Methane* **2**, 440-451. DOI:10.3390/methane2040030.



A graph depicting methane flux measurements. [Courtesy Lawrence Berkeley National Laboratory]



A graph showing methane emission increasing over time as measured by field chambers. Reprinted with permission from Elsevier from Villa, J. A., et al. 2021. "Ebullition Dominates Methane Fluxes from the Water Surface Across Different Ecohydrological patches in a Temperate Freshwater Marsh at the End of the Growing Season," *Science of the Total Environment* **767**, 14498. DOI:10.1016/j.scitotenv.2020.144498.

***Bathyarchaeota* subgroups and operational taxonomic units display phylogenetically conserved abundance patterns in the Old Woman Creek wetland, Ohio, correlating to geochemical measures.** Reprinted with permission from Wiley from Narrowe, A. B., et al. 2017. "High-Resolution Sequencing Reveals Unexplored Archaeal Diversity in Freshwater Wetland Soils," *Environmental Microbiology* **19**(6), 2192-2209. DOI:10.1111/1462-2920.13703.

Genetic sequence data. [Courtesy Adobe Stock]

Appendix F

Acronyms and Abbreviations

16S	16S ribosomal RNA sequencing
AI	artificial intelligence
AI4CH$_4$	Artificial Intelligence for the Methane Cycle workshop
AI4ESP	Artificial Intelligence for Earth System Predictability workshop series
ANN	artificial neural network
API	application programming interface
ARM	Atmospheric Radiation Measurement user facility
ASCR	DOE Advanced Scientific Computing Research program
BER	DOE Biological and Environmental Research program
BETO	DOE Bioenergy Technologies Office
BSSD	BER Biological Systems Science Division
CH$_4$	methane
CLM4Me	Community Land Model, version 4.0, for methane
CMIP5	Coupled Model Intercomparison Project Phase 5
CMIP6	Coupled Model Intercomparison Project Phase 6
CNN	convolutional neural network
CO$_2$	carbon dioxide
COMPASS	Coastal Observations, Mechanisms, and Predictions Across Systems and Scales
COSORE	Continuous Soil Respiration database
CPU	central processing unit
DBTL	design-build-test-learn cycle
DNN	deep neural network
DOE	U.S. Department of Energy
DOI	digital object identifier
E3SM	Energy Exascale Earth System Model
EESSD	BER Earth and Environmental Systems Sciences Division
ELM	E3SM's Land Model
ESnet	ASCR Energy Sciences Network
ESS-DIVE	Environmental System Science Data Infrastructure for a Virtual Ecosystem
FAIR	findability, accessibility, interoperability, and reusability
FLIR	forward-looking infrared
GAN	generative adversarial network
GCAM	Global Change Assessment Model
GEM	genome-enabled model
GSA	global sensitivity analysis
HPC	high-performance computing
ICON	Integrated Coordinated Open Networked science principles
ILAMB	International Land Model Benchmarking
IMEO	United Nations Environment Programme's International Methane Emissions Observatory
IPCC	Intergovernmental Panel on Climate Change

KBase	DOE Systems Biology Knowledgebase
LLM	large language model
LSTM	long short-term memory network
MCMC	Markov Chain Monte Carlo Bayesian technique
MIGS	minimum information about a genome sequence
MIMARKS	minimum information about a marker gene sequence
MIMS	minimum information about a metagenome sequence
MIxS	minimum information about (X) any sequence
ML	machine learning
ModEx	model-experiment framework
MONet	Molecular Observation Network
MRV	monitoring, reporting, and verification
MTE	model tree ensemble
NEON	National Ecological Observatory Network
NGEE	Next-Generation Ecosystem Experiments
NMDC	National Microbiome Data Collaborative
NN	neural network
NSF	National Science Foundation
PFLOTRAN	Massively Parallel Reactive Flow and Transport Model for Describing Subsurface Processes
QA	quality assurance
QC	quality control
SciDAC	DOE Scientific Discovery through Advanced Computing program
SOC	soil organic carbon
SPRUCE	Spruce and Peatland Reponses Under Changing Environments
Tg	teragram
TPU	tensor processing unit
TROPOMI	TROPOspheric Monitoring Instrument
UAV	unmanned aerial vehicle
UQ	uncertainty quantification