Process Discovery through Assimilation of Complex Biogeochemical Datasets

M. Zavarin (LLNL), H. Wainwright (LBNL), P. Nico (LBNL), C. Steefel (LBNL), C. Varadharajan (LBNL), J.R. Bargar (SSRL), J. Damerow (LBNL), N.D. Ward (PNNL), V.L. Bailey (PNNL), K.M. Kemner (ANL)

Focal Area(s)

This white paper addresses two of three focal areas identified in the white paper call: 1) Biogeochemical data acquisition and assimilation enabled by machine learning and 3) Insight gleaned from complex data using AI. We focus on AI application to complex biogeochemistry (BGC) data (e.g. laboratory experimental data, field manipulation data, literature data), which is an untapped source of information for improving Earth System Predictability (ESP).

Science Challenge

Laboratory experiments and field manipulation experiments are critical to interrogate the impact of hydrological and climate perturbations on BGC processes in a controlled environment. Hydrologically-driven BGC processes are a key aspect of ESP, particularly at dynamic interfaces (e.g. terrestrial-aquatic interfaces, hot spots/hot moments), because BGC processes govern the cycling of nutrients, metals, and organic matter. However, BGC experiments yield a complex array of data types (spatiotemporal observational and laboratory measurements, microscopy image data, spectroscopy data, etc.), which hampers assimilation and analysis, as well as the application of machine learning (ML). Ensuring that these data are findable, accessible, interoperable, and reusable (FAIR) is of paramount importance. AI/ML methods are poised to transform the way we incorporate complex BGC laboratory and field manipulation experimental data into earth and environmental systems models, quantify data uncertainty, design future experiments, and develop new models with unprecedented fidelity and resolution.[1]

Rationale

There is a paradigm shift required to make complex BGC data in compliance with the FAIR principles, increase fidelity of Earth System models, and embrace the MODEX paradigm across scales. The paradigm shift requires: **(1) assimilation of FAIR laboratory and field manipulation experimental data; and (2) application of ML to these data to incorporate findings and understanding into a variety of process and predictive models that span a wide range of spatiotemporal scales.** Its success will open opportunities to develop new ML algorithms such as ML-trained surrogate models and uncertainty quantification (UQ) to transform ESP and integrate knowledge across scales.

To effectively use information from these complex experimental datasets, strong data management practices including quality control, uncertainty characterization, and use of community (meta)data reporting formats that enable data discovery, interpretation, and integration are critical. Scalable open-source software for advanced analysis techniques such as data mining, machine learning, pattern scaling, and visualization will enable mechanistic scientific discovery as well as development of new predictive capabilities. Consistency across models of different scales and complexity will enable scientists to more easily identify unknown coupled BGC processes, inform simpler models with more complex ones, add higher resolution or more detailed process understanding for simulations of regions or phenomena, and couple models across disciplines as needed to address critical science questions.[3]

DOI: https://doi.org/10.2172/1769767
DOE BER, AI4ESP 2021 #AI4ESP1152 (https://www.ai4esp.org)

Process Discovery through Assimilation of Complex Biogeochemical Datasets

An important component of EESSD's effort involves promoting effective data management, which includes developing community data standards and formats and sharing and preserving data, to increase the pace of scientific discovery and ensure scientific integrity. The goal of the ESS Data Infrastructure for a Virtual Ecosystem (ESS-DIVE, https://ess-dive.lbl.gov/) data archive is to publish and preserve DOE's diverse Earth and environmental science data so that it can be discovered, integrated, and reused by the scientific community. *The development of AI approaches to assimilate and interpret these BGC data, develop process-level understanding of complex systems, develop scaling approaches, and integrate with watershed and Earth system models will open new opportunities to improve ESP.*

Narrative

The EESSD vision is to develop an improved capability for ESP on seasonal to multidecadal time scales to inform the development of resilient U.S. energy strategies. This vision is supported by laboratory experiments, long-term field experiments, DOE user facilities, modeling and simulation, uncertainty characterization, best-in-class computing, process research, and data analytics and management.[3] Laboratory and field experiments are particularly important for the future prediction under climate change, facing more frequent extreme events associated with the water cycle, in which the inference based on historical data often fails due to the lack of representations of extreme events and conditions. Laboratory and field manipulation experiments can create extreme conditions or new normal conditions in a controlled environment, which provide fundamental understanding as well as extending the observational range beyond the natural environment. Despite such importance, the application of ML to these data has been limited. Small-dimension experimental data collected in laboratory settings are fundamentally different from large formatted remotely sensed information that is much more amenable to "big data" approaches available to modern data science.

We propose a paradigm shift to integrate laboratory and field manipulation experiments and their data into the Earth Systems models, to account for uncertainties from raw data to model simulations as well as to extend into the design of experiments based on model simulations. In addition, we aim to establish the framework to facilitate the feedback from modeling to lab-based and field-based experiments, which improves the last link in the MODEX cycle. To achieve this paradigm shift, we envision that four ML topics will play a critical role: **(1) natural language processing for data discovery from existing diverse literature, (2) systematic methods to compare multiple models with multiple/global datasets to account conceptual/structural uncertainties, (3) surrogate models or oracle-based approaches to optimize experimental design, (4) federated experiments across multiple labs and institutions.**

Experimental studies and mechanistic model development happen on a course of multiple generations of scientists. Although systematic data curation and management is currently possible, integrating historical data (i.e., data mining) must also be considered, especially for costly laboratory and field manipulation experiments. The potential value of historical data has been widely recognized.[4,5] As an example, in the geosciences, a manual data mining effort focused on mineral-water interface data for the mineral hydrous ferric oxide[6]– yielded a robust database that has been applied widely throughout the geochemistry community (4916 google scholar citations). Similarly, the active data mining effort of RES^{3}T (https://www.hzdr.de/db/RES3T.login) contains 3172 references and includes reactions between

147 minerals and 148 ligands and a total of 7062 reaction constants that span across all known surface complexation models. ML-empowered **natural language processing and automated data discovery/extraction from diverse literature** – increasingly used in chemical and material sciences[4,5,7-11] – can similarly transform data utilization in the Earth sciences to improve ESP.

A formal benchmarking and model comparison framework has been increasingly used for climate models and reactive transport models across the EESSD community (e.g., iLAMB).[14,15] However, these benchmark exercises primarily start from given parameters; without considering the uncertainties associated with raw datasets and fundamental geochemical models. The framework to directly integrate BGC data and/or databases with models will enhance the ability to account for uncertainties. In addition to traditional approaches such as AIC and BIC, new machine learning approaches such as multi-model hierarchical sensitivity analysis framework have a great potential to accelerate model benchmarking and intercomparison as well as inform data collection/generation; for example, which parameters are most important to reduce model uncertainties.[12]

In parallel, a recent key innovation of ML in chemistry and material sciences has been focused on ***data-driven design***. Rather than considering many combinations of multiple experimental parameters and doing costly experimental measurements, ML is used to develop **high-capacity regression or surrogate models** – ***oracles*** – trained on labeled data, which can be leveraged in an in silico search for optimal parameters that provide desirable material/chemical properties.[13] The adaptive strategy to include new datasets and update the regression models – referred as ***autofocusing*** – has also been implemented.[13] Although ESP does not mean to create any particular outcome, we may take advantage of these ML methods to find the particular parameter space that creates particular responses in BGC. In particular, BGC is known to exhibit nonlinear behaviors such as hot spots and hot moments; focusing the parameter space to facilitate interrogation of underlying specific reactions, rate controls, and process interactions would be highly beneficial. In addition, ESP yields unique problems that have not been explored in general chemical and material sciences: can we change the objective function to minimize the uncertainty rather than finding a particular space or can we implement surrogate models or oracle based mechanistic models in the design algorithm rather than data-driven regression models? If implemented, **AI can help design BGC experiments by identifying the parameter ranges needed to minimize uncertainty in the predictions.** Finally, this framework – comprehensive databases, data-model integration, and model-based design – will open the door to experimental innovations in the national lab network. Once particular parameter space and responses are identified through comprehensive database and data-model integration efforts, **we may federate laboratory experiments into multiple institutions so that we can cover the parameter space effectively and generate large numbers of data that can support ML and ESP**. BER's ESS-DIVE can play a key role in promoting use of consistent reporting formats for datasets that enable effective data discovery, integration, and reuse. AI systems can then be employed to automate the creation of FAIR BGC data, integrate it into knowledge repositories, and provide the architectural basis for new data infrastructure necessary to accelerate AI training and model development.

Process Discovery through Assimilation of Complex Biogeochemical Datasets

Suggested Partners/Experts (Optional)

Yong Han (LLNL) - AI applications to natural language processing and materials science applications
Jason Hou (PNNL) - Machine Learning, Data Mining, Uncertainty Quantification, Extreme Events, Complex Systems

References

1. Department of Energy, 2020, AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science.
2. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, https://doi.org/10.1038/sdata.2016.18
3. Department of Energy, 2018, Earth and Environmental Systems Science Division: Strategic Plan, DOE/SC–0192.
4. Batra, R. 2021, Accurate machine learning in materials science facilitated by using diverse data sources. Nature. https://doi.org/10.1038/d41586-020-03259-4
5. Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y., and Hiszpanski, A.M. 2020. Data-driven materials research enabled by natural language processing and information extraction, Appl. Phys. Rev., 7:041317, doi: 10.1063/5.0021106.
6. Dzombak, D.A. and Morel, F.M.M. (1990) Surface complexation modeling : hydrous ferric oxide. Wiley, New York.
7. Chen, C., Zuo, Y., Ye, W. *et al.* 2021. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat Comput Sci* **1,** 46–53. https://doi.org/10.1038/s43588-020-00002-x.
8. Swain, M.C., and Cole, J.M. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature Journal of Chemical Information and Modeling 56(10):1894-1904, DOI: 10.1021/acs.jcim.6b00207
9. Mukaddem, K.T., Beard, E.J.,Yildirim, B., and Cole, J.M. 2020. ImageDataExtractor: A Tool To Extract and Quantify Data from Microscopy Images, Journal of Chemical Information and Modeling 60(5):2492-2509, DOI: 10.1021/acs.jcim.9b00734
10. Kim, H., Han, J. and Han, T.Y. 2020. Machine vision-driven automatic recognition of particle size and morphology in SEM images, Nanoscale, 12:19461.
11. Hiszpanski, A.M., Gallagher, B., Chellappan, K., Li, P., Liu, S., Kim, H., Han, J., Kailkhura, B., Buttler, D.J., and Han T.Y. 2020. Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge, J. Chem. Inf. Model., 60:2876−2887.
12. Ju, J., Dai, H., Wu, C., Hu, B. X., Ye, M., Chen, X., Gui, D., Liu, H., & Zhang, J. (2021). Quantifying the Uncertainty of the Future Hydrological Impacts of Climate Change: Comparative Analysis of an Advanced Hierarchical Sensitivity in Humid and Semiarid Basins, *Journal of Hydrometeorology*, https://doi.org/10.1175/JHM-D-20-0016.1
13. Fannjiang, C., & Listgarten, J. (2020). Autofocused oracles for model-based design. arXiv preprint arXiv:2006.08052.
14. Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., ... & Randerson, J. T. (2018). The International Land Model Benchmarking (ILAMB) system:

design, theory, and implementation. Journal of Advances in Modeling Earth Systems, 10(11), 2731-2754.

15. Steefel, C. I., Yabusaki, S. B., & Mayer, K. U. (2015). Reactive transport benchmarks for subsurface environmental simulation. Computational Geosciences, 19(3), 439.