AI to Automate ModEx for Optimal Predictive Improvement and Scientific Discovery

Authors: Shawn P. Serbin (BNL), Scott E. Giangrande (BNL), Chongai Kuang (BNL), Nathan Urban (BNL), Line Pouchard (BNL)

Focal Areas

- Data acquisition and assimilation enabled by machine learning, AI, advanced experimental optimization, unsupervised learning, and hardware-related AI efforts.
- Predictive modeling through AI techniques and AI-derived model components; Using AI to design a hierarchical model prediction system consisting and model selection.
- Interrogating complex data (observed and simulated) using AI, big data analytics, and other advanced methods such as explainable AI and *physics- or knowledge-guided* AI.

Science Challenge

Our capacity to predict the responses of Earth's ecosystems to climate change lags behind the ability to monitor and measure changes in the biosphere. Addressing this challenge requires a transformational shift in the computational and data infrastructures used to inform climate predictions, focused on the development of accessible, scalable, and transparent tools that integrate the expertise of modelers, domain experts, and empiricists/observationalists to accelerate the pace of discovery (Fer et al., 2021). While climate-relevant datasets span critical observational and process scales, the chief roadblock to Earth System Model (ESM) improvement is the mismatch between the breadth of available data and the limited capacity to use this information to inform our process understanding of the atmosphere, hydrologic cycle, and land-atmosphere feedbacks.

The upcoming deployment of the ARM Mobile Facility 3 (AMF3) to the Southeast U.S. (SEUS, expected start, 2023) offers a unique set of challenges and opportunities for developing a cyberinfrastructure and optimization-driven Model-Experimental (ModEx) testbed framework leveraging AI/ML tools. This framework can evaluate end-to-end solutions that guide data acquisition and optimal design strategies, implement complex multiscale and multi-disciplinary data synthesis and prediction tools, and promote model-data integration, hypothesis testing, and uncertainty quantification (UQ) in a region of the US characterized by strong land-atmosphere coupling. The primary outcomes of this framework would be an ability to:

- Optimize data collection strategies and syntheses, also applicable to campaign design;
- Identify critical predictive uncertainties and observational requirements;
- Advance analysis-ready, scale-aware datasets for discovery science and modeling;
- Accelerate data ingest into process or hybrid process/ML modeling frameworks.

Rationale

An accurate process understanding of the global water and energy cycles is essential for advancing climate prediction capabilities, and our ability to anticipate the impacts of extreme weather and drought on (i) ecosystems and society, (ii) food security, and (iii) resilient transportation and energy infrastructures. Hydrologic cycle challenges encompass a range of scientific disciplines, yet are rarely investigated in a coordinated way, leading to inconsistencies in model-process

DOI: https://doi.org/10.2172/1769662
DOE BER, AI4ESP 2021 #AI4ESP1119 (https://www.ai4esp.org)

fidelity and complexity. The upcoming AMF3 SEUS deployment seeks to formally connect land-surface and atmospheric processes over a region characterized by complex surface and atmospheric states, and home to strong land-atmosphere water cycle coupling, frequent clouds and routine extreme precipitation (e.g. Feng et al. 2018). The SEUS represents a timely and unique opportunity to develop a novel AI-guided, optimization-driven ModEx testbed to address relevant water cycle challenges laid out in the EESSD's Strategic Plan (U.S. DOE, 2018) and the predictive uncertainties related to the hydrologic cycle and future vulnerabilities.

Narrative

An AI/ML ModEx approach together with comprehensive data and modeling cyberinfrastructure is critical to efficiently integrate observations, link models, and disciplines to transform traditional DOE campaigns or synthesize multiple multi-scaled efforts. SEUS is a relevant starting point for impactful US-based water cycle discussions, and a unique opportunity for an initial testbed to enhance an upcoming DOE campaign through: (i) AI/ML-guided ModEx for design or evaluation of data acquisition strategies, (ii) model uncertainty quantification (UQ), (iii) guiding observational requirements for multiscale modeling, and/or (iv) assimilation to address and reduce uncertainty in ecohydrology and coupled land-surface atmospheric modeling. This framework could be applied more broadly to past, present, and future efforts (e.g., SGP, Ameriflux, NGEEs). AMF3 SEUS is also ideal for evaluating DOE and related technologies in support of a larger comprehensive AI/ML infrastructure, including edge-computing informed by domain experts and model UQ to inform deployment of sophisticated networks of 5g-enabled sensors to capture SEUS surface heterogeneity and address relevant water cycle science drivers.

Theme 1. AI-guided Data Acquisition Strategies and Optimal Design:

DOE campaigns collect a wide range of datasets, including high-frequency land-surface, cloud and aerosol measurements, often resulting in a significant data volume. For example, current DOE water cycle activities rely heavily on resource-intensive radar measurements for cloud process studies that may not efficiently target those key cloud processes with existing operational modes, or be optimally sited. End-to-end ModEx simulations, leveraging AI tools including ML model surrogates, and reinforcement-learning, can be used to develop optimal sampling for AMF3 SEUS and other campaigns, addressing key challenges associated with operations design. Similarly, these methods can iteratively adapt sampling strategies or define optimal measurement frequencies over time (seasonally or annually) by proposing new collection or enhanced deployment strategies that target specific improvements in model performance, together with hybrid or emulator model UQ. Similarly, AI/ML edge-computing for rapid data QA/QC, pattern recognition, and/or AI-assisted dimensionality reduction could help inform adaptive atmospheric measurement strategies, reduce data volume by rapidly targeting data collection, and/or separating observations from background conditions. This illustrates an added pathway wherein model needs are used to iteratively inform measurement requirements. DOE campaigns like SEUS are also ideal testbeds to advance AI-assisted 5G "smart sensor" networks targeting key variables (e.g. micromet, PM2.5/10) and distributed methods for integrating data across multiple facilities.

Theme 2. AI-assisted Data Curation for Discovery:

DOE observations are typically multidisciplinary, diverse, and spanning multiple spatio-temporal scales. This includes the expected AMF3 SEUS datastreams, well-suited for the application of a modern AI/ML cyberinfrastructure and data analytics tools to efficiently curate, harmonize, and synthesize diverse, often noisy or of varying quality, datasets into analysis-ready, scale-aware observations suitable for multi-scale modeling. Guided by FAIR principles, Findability, Accessibility, Interoperability, and Reusability (Wilkinsons et al. 2016), this novel infrastructure could provide easy access to datasets sourced from multiple archives (e.g. Ameriflux, DAAC, ESS-DIVE) to serve a wide-range of scientific applications (e.g. Pouchard et al. 2013), based on community standards (e.g. Ely et al 2021). Given the wealth of DOE data, accessible, harmonized, trustworthy, and QA/QC datasets (with quantitative uncertainties) are essential for future broad-scale applications of AI/ML and ModEx. A DOE SEUS testbed would evaluate modern data processing tools, including globally-persistent unique identifiers to facilitate provenance tracking across several archives, federated data searches, and maintaining standards compliance. Data curation informed by AI can identify and/or gap-fill missing or erroneous data through clustering across variables and scales. AI's ability to process large data volumes enables the discovery of functional relationships between variables (e.g. temperature and evapotranspiration), useful for predictive modeling and benchmarking (Fer et al., 2021). Integration of ontology into metadata will support discoverability in persistent data pipelines and workflows, and ensure interoperability across data and analysis (e.g. Python, R) platforms. Supporting end-to-end ModEx, an AI data testbed would facilitate hyperparameter sweeps together with the tracking of simulations through metadata acquisition to compare ML/process model results and automate model retraining or updated simulations, given new data.

Theme 3. Hybrid ML/process-model UQ and Assimilation:

A comprehensive AI-guided end-to-end ModEx framework requires the ability to synthesize and integrate observations to inform our predictive understanding of ecosystems and climate, address uncertainties and advise observation requirements, or test competing hypotheses. For DOE campaigns such as SEUS, the dependence on multiscale data and models necessitates computationally efficient AI/ML tools to manage information flows and provide tractable approaches for model UQ and assimilation to inform predictions, feeding back to data needs. The ModEx framework can leverage AI, including hybrid modeling methods that replace select underlying physical process representations or scales with an ML approximation. In addition, physical model emulators, tuned from AI-optimized ensembles, can be used as a replacement for the full coupled model framework. Simplifying the computational costs with AI allows for the interrogation of the system to identify the information contribution of data, identify the most uncertain processes or scales, target new datasets, assimilate observations, and evaluate and test different competing hypotheses. A SEUS application could use AI to address challenges related to informing land-surface processes (e.g. surface energy balance, turbulence) and their connection with larger atmospheric impacts (two-way interactions), a core focus of the AMF3 campaign.

References

Ely, K. S., A. Rogers, D. A. Agarwal, E. A. Ainsworth, L. P. Albert, A. Ali, J. Anderson, M. J. Aspinwall, C. Bellasio, C. Bernacchi, et al. (2021). A reporting format for leaf-level gas exchange data and metadata. Ecological Informatics **61**, 101232. DOI: https://doi.org/10.1016/j.ecoinf.2021.101232

Feng, Z., Leung, L. R., Houze, R. A., Hagos, S., Hardin, J., Yang, Q., et al. (2018). Structure and evolution of mesoscale convective systems: Sensitivity to cloud microphysics in convection-permitting simulations over the United States. *Journal of Advances in Modeling Earth Systems*, 10, 1470– 1494. https://doi.org/10.1029/2018MS001305

Fer, I., A. K. Gardella, A. N. Shiklomanov, E. E. Campbell, E. M. Cowdery, M. G. De Kauwe, A. Desai, M. J. Duveneck, J. B. Fisher, K. D. Haynes, F. M. Hoffman, M. R. Johnston, R. Kooper, D. S. LeBauer, J. Mantooth, W. J. Parton, B. Poulter, T. Quaife, A. Raiho, K. Schaefer, S. P. Serbin, J. Simkins, K. R. Wilcox, T. Viskari, and M. C. Dietze. (2021). Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration. Global Change Biology **27**, 13-26. DOI: https://doi.org/10.1111/gcb.15409

Pouchard, L. C., M. L. Branstetter, R. B. Cook, R. Devarakonda, J. Green, G. Palanisamy, P. Alexander, and N. F. Noy. 2013. A Linked Science investigation: enhancing climate change data discovery with semantic technologies. Earth Science Informatics **6**, 175-185. DOI: 10.1007/s12145-013-0118-2

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, https://science.osti.gov/-/media/ber/pdf/workshop-reports/2018_CESD_Strategic_Plan.pdf

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018. DOI: 10.1038/sdata.2016.18