Learning from learning machines

improving the predictive power of energy-water-land nexus models with insights from complex measured and simulated data

Authors

James B. Brown, Michael Sohn, Utkarsh Mital, Dipankar Dwivedi, Haruko Wainwright, Carl Steefel, Eoin L Brodie, William Collins, Daniel A. Jacobson, Michael W Mahoney, Tianzhen Hong, Christoph Gehbauer, Doug Black, Thomas Kirchstetter, Daniel Arnold, Sean Peisert

Focal Area(s)

- Insights gleaned from complex data (observed and simulated) using AI
- Predictive modeling through the use of AI techniques and AI-derived model components, including physics- and knowledge-informed models
- Energy-water-land nexus and integrated energy systems – models of MultiSector dynamics

Science Challenge

Scientific communities are in need of tools for the computational integration of physics-based models, experimental data, and empirical/observational studies across a broad range of temporal and spatial scales to explore and meet EESSD grand challenges. We require AI algorithms for the discovery of process-drivers in the Earth-energy-human system and to link them with complex measured and simulated data. We aim to understand the energy-water-land nexus, particularly under extreme forcing scenarios and rare events, leveraging scale-aware AI process models, including probabilistic uncertainties, and benefiting from the emerging 5G-enabled landscape-scale sensing and edge computing capabilities. These models are needed to identify instabilities and tipping points that manifest extreme system behaviors with consequences for integrated energy systems and the environment. This work requires fundamental advances in uncertainty quantification, in particular, to identify and model unlikely but catastrophic outliers. Specifically, *we need models that are interpretable to domain scientists and, essentially, explainable to public and private stake holders*.

Rationale

Interactions between human and environmental systems are intrinsically complex and massively multi-scale: extreme weather events result in correspondingly extreme demands on the power grid, and drought intensification alters water quality as well as availability – which comes with power requirements. These events in turn can have broad deleterious effects on critical infrastructure and ecosystem services. Extreme weather events (e.g., heatwaves, wildfires) are becoming more frequent and intense, leading to multi-billion-dollar-scale economic losses and loss of life and resources. Quantitative understanding of such events is limited, and predictive understanding is nonexistent. This is due, in large part, to limits in our capacity to learn from complex data – to discover high-order interactions, and understand rare events and to model the likelihood of extrema. Measured data on extrema is, by definition, rare – running models far from the conditions under which they were conceived and fitted creates onerous challenges, and

DOI: https://doi.org/10.2172/1769736
DOE BER, AI4ESP 2021 #AI4ESP1019 (https://www.ai4esp.org)

quantifying uncertainty and effect size is, even at exascale, infeasible without advances in applied mathematics. This is due to a fundamental property of data-driven models: models are unreliable far from the support on which they are trained. However, first principles models that capture physical properties of the world extrapolate well. Our challenge is to develop methods to learn from AI models about the world – not just about the data on which they are fitted. Central to this aim is developing new inferential and uncertainty analysis procedures for extrema (outliers). Learning to direct risk-averse and risk-taking decisions for integrated energy infrastructure in environmental contexts is a grand challenge. The development of AI to learn the dynamic, multi-level interdependencies between MultiSector human systems, the built environment, energy capacity, and natural environments stands to transform our ability to make educated decisions in our stewardship of our shared, multigenerational resources. Success will enhance our understanding of the impacts of complex stressors on human systems and infrastructure, particularly vulnerabilities and risks at the energy-water-land nexus. The developed AI toolkit will also support U.S. cities' energy and environmental goals (e.g., 100% clean energy, climate neutrality).

Barriers to progress: we need a new generation of AI technology to meet EESSD needs. AI, as it exists today, is not sufficient to provide transformative scientific advances. We need to drive the development of AI models bespoke for EESSD grand challenges if we expect to achieve breakthroughs in 2-5 years. This research will be especially useful to model process extrema that are, at present, poorly supported due to the limited integration of physics-like first-principles models. We also know that data gathering and processing of unaccounted amounts of information from integrated energy systems (e.g., electrical grid, buildings, transportation) and water-energy infrastructure is infeasible using classical computational methods within time scales for systems management and optimization, or for cybersecurity.

What is the significance to "extreme" water cycles? To understand extreme water cycles, we need new hypotheses – extant knowledge and models are not sufficient. AI for hypothesis discovery and counterfactual reasoning will require bespoke methods development. New methods will enable hypothesis discovery from the integration of regional scale remote sensing, ground-based data, and multi-domain models. These hypotheses must be grounded with existing, though limited, data from energy-water-land nexus systems.

Narrative

Hypothesis discovery from data – learning mechanisms from observations. Recall that an AI model can be viewed as a function f that takes as input some data x and returns as output a response y.Then, f can be viewed as a composition of two functions, $y = f(x)$ with $f = h \circ g$ and $g : M \longrightarrow N$, (two topological spaces) where f is the overall AI model, g computes the model's data representation, h computes the response, and the composition $h \circ g$ indicates that $g(\cdot)$ is applied to x, and then $h(\cdot)$ is applied to $g(x)$. Often, but not always, it holds that $\dim(N) \ll \dim(M)$. In this way, g encodes the patterns learned by the AI model, and h encodes the map that links these patterns to outcomes in the world,

y. For example, with a Feed-Forward Neural Network, it may be useful to choose g to consist of all layers except the last decision layer, or just initial encoding layers. We will study representations discovered by g, as well as their meaning, encoded by h, in terms of that are useful and familiar to Earth and Energy systems scientists. We will use techniques from nonparametric statistics to study uncertainty in the latent space N. Importantly, we will not pursue arbitrary metric embeddings, but rather the natural geodesics imposed by h on g – this distinction is essential to compute meaningful measures of confidence and uncertainty. A testable hypothesis discovered from analysis of data is then of the form:

$$y_0 = h\big(g(x_0)\big) \mid x_0 \in U_0 \subset \mathfrak{X}$$

where U_0 is a sparse support in the domain $\mathfrak{X}$ of g (the input space). Note that if $\mathfrak{X}$ lacks explicit semantic meaning, e.g., imaging data, then sparsity is asserted in the latent space, the domain of h rather than g. We will develop uncertainty quantification and propagation procedures for AI models. Promising work on risk-controlling prediction sets points us toward general strategies for uncertainty propagation through arbitrary models – AI or otherwise. Success will endow AI models with the same interpretability and capacity for hypothesis discovery as statistical models.

Physics and knowledge informed learning. We've described methods to extract knowledge from AI – now we turn to the problem of *describing knowledge to AI*. For many processes, e.g., reactive transport, power transmission, we have excellent process models, and some idea of how they can be coupled across scales. For others, e.g., carbon cycling, we know little about the drivers of terrestrial carbon stores and their residency-times – or how integrated energy systems may be used to minimize anthropogenic carbon emissions. Where components of predictive problems can be constrained by physics, there is a rich literature upon which to draw. However, when spatiotemporal scales of measurements and processes are mismatched, as is often the case, e.g., in metabolic processes or MultiSector dynamics, conservation laws fail, and explicit physics are not available. However, it is often still possible to encode prior knowledge, e.g., in the form of knowledge graphs, into AI models. We will develop techniques using reinforcement learning to perform knowledge informed and knowledge constrained learning.

Next-generation Earth-energy-human systems models. We aim to integrate AI models with current and future components of the Energy Exascale Earth System Model (E3SM). In the Coastal Observations, Mechanisms, and Predictions Across Systems and Scales (COMPASS) pilot study, we are working to identify process drivers of ecosystem transitions between metastable equilibria, e.g., from carbon sequestering to carbon emitting. We will study integrated energy system models that couple across scales (e.g., from buildings to grids) and domains (e.g., transportation and fuel, natural systems), and will use HPC-enabled simulations to provide a "playground" for AI models using reinforcement learning. We will identify emergent parameterizations that enable the integrated control of complex systems.

Summary. AI stands to revolutionize BER science. The revolution won't happen without input into methods development from BER researchers.

Learning from learning machines

improving the predictive power of energy-water-land nexus models with insights from complex measured and simulated data

Suggested Partners/Experts: Peter Bickel (Berkeley), Bin Yu (Berkeley), Claire Tomlin (Berkeley), Anil Aswani (Berkeley), Juli Muller and Andrew Jones (LBNL), Ghanshyam Pilania (LANL), Sumanta Basu (Cornell), Nancy Zhang (UPenn), Maya Gupta (Google), Maziar Raissi (U. Colorado), Kyle Cranmer (NYU), Mike Kirby (Utah), Nathan Kutz (Washington), Cynthia Rudin (Duke), Mark Anderson (Pattern Computer), Ken Kreutz-Delgado (UCSD)