AI Automated Discovery of New Climate Water System Knowledge from Models and Observations

André Goncalves[1], Donald Lucas[2]
1. andre@llnl.gov, Computational Engineering Division, LLNL
2. ddlucas@llnl.gov, Atmospheric, Earth, and Energy Division, LLNL

Primary Focus Area: This paper addresses focus area 3, "Insight gleaned from complex data."

Science Challenge: The climate-water system is highly complex, containing a multitude of positive and negative feedbacks, time scales stretching many orders-of-magnitudes, and non-linear, connected processes. Powerful artificial intelligence (AI) methods can revolutionize and automate the discovery of new knowledge and relationships in the climate-water system, which will improve understanding and predictability of extreme hydrological events.

Rationale: Discovery of new feedbacks, teleconnections and causal relationships in the climate system can improve our understanding of climate change and lead to better, more predictive climate models. However, progress in climate science is hindered by the long, painstaking steps required to extract new information from complex systems. Climate scientists have spent decades studying the Madden Julian and El Nino Southern Oscillations that drive the water cycle at intra-seasonal to inter-annual timescales, but explanatory mechanisms of the coupled ocean-atmosphere processes needed to better predict and forecast these phenomena remain elusive. We envision a future AI system that accelerates learning and discovery of new causal relationships in the climate system. This AI system will take in a climate data from observational systems and model simulations, assimilate and process this data, and automatically generate plausible mechanisms and explanations for relationships in the data. The proposed AI framework is referred to as CHANGE: Climate Hypothesis ANalysis and GEneration.

Narrative: The traditional scientific process of discovering new information is time-consuming and arduous. Scientists formulate a hypothesis, gather observations, and conduct analyses to determine if the hypothesis is supported or can be discounted. This global endeavor requires years or decades of dedicated teams of researchers from multiple disciplines, large funding amounts and strategic coordination.

Starting with climate observational data, one important goal is to learn new underlying biogeochemical and physical relationships that occur in the climate-water system but are not represented in climate models. Recent progress in artificial intelligence and machine learning has opened potential new doors to discover new processes, relationships, and physical laws in the climate system. Given noisy data sampled from simplified fluid systems, researchers have, for example, extracted the Navier Stokes and shallow water equations [Zhang and Lin, 2018].

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.

DOI: https://doi.org/10.2172/1769658
DOE BER, AI4ESP 2021 #AI4ESP1051 (https://www.ai4esp.org)

We propose CHANGE, an AI-powered framework for automated hypothesis generation and testing for water-cycle systems. The framework is composed of two complementary phases: 1) hypothesis generation from climate observations; and 2) hypothesis testing based on climate model data. Figure 1 shows a schematic view of the proposed approach.

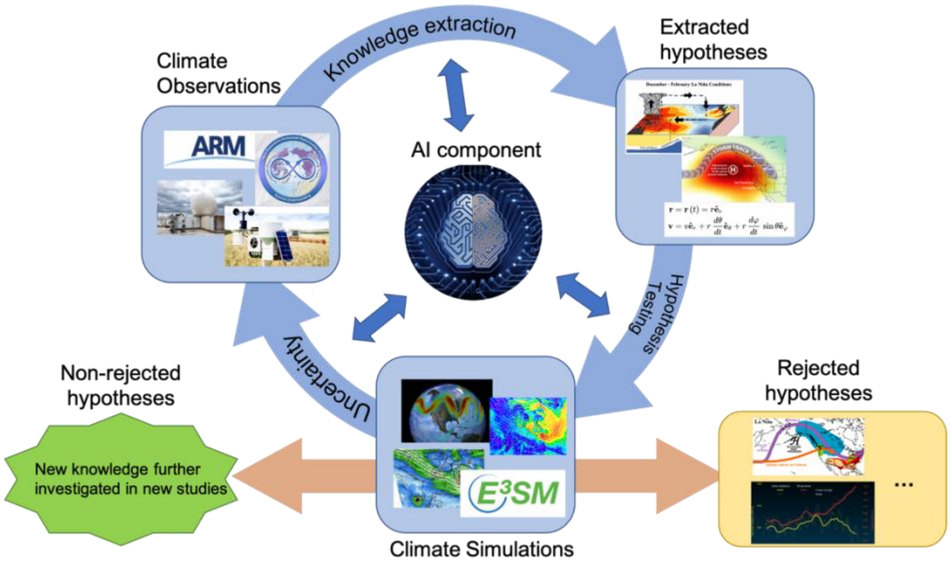

Figure 1: Scientific hypothesis generation and testing for new knowledge discovery with artificial intelligence.

In the hypothesis generation phase, the AI component learns concise, yet powerful latent representations of the sub-systems present in climate-water systems via observation data. These representations are then related to each other via a hierarchical structure. For this process, instead of relying on human-provided annotations, the AI component creates labels from the data itself by exposing relationships between the data's parts. It basically provides its own supervision. A few AI paradigms can be used to extract computational representations of the water-climate dynamical systems including *unsupervised learning*, *self-supervised learning*, *generative adversarial learning* and *reinforcement learning* techniques. All of these approaches do not require human intervention for their operation, as explicit labels are not necessary. Initial studies for language understanding applications [Devlin *et al.*, 2019] the aforementioned learning paradigms have shown that it was capable of learning high-level language concepts and how they related to each other. These methods were able to unveil language rules only by looking at a collection of texts, without any human intervention or annotation. Similarly, we expect that given enough computational power and data, the AI system can also uncover important processes that drive climate water systems.

For the hypothesis testing phase, a probabilistic predictive AI component is responsible for testing whether the generated hypotheses can be falsified based on climate model data collected from existing sources. To each hypothesis, the AI component provides two outcomes: a falsifiability score and a level of uncertainty about its own prediction. A given hypothesis can be confronted to any climate model, via their outputs. If most of the climate models reject the hypothesis, it can possibly be an unsupported hypothesis and might be discarded. In case many climate models cannot falsify the hypothesis, it might suggest a valid hypothesis. The outcomes of the AI system can also provide information for the developers of the climate models for which the hypothesis was falsified, as they may be missing important climatological processes present in the observation data. This particular component could be implemented as a Bayesian graph neural network [Hasanzadeh *et al.*, 2020], which is capable of providing both predictions and uncertainties. The input is hypothesis that is represented as a hierarchical structure of the concepts and their relations. The output is the likelihood of being falsified based on climate model data, and uncertainty measurement about its prediction.

The outcomes of the hypothesis testing: falsifiability prediction and uncertainty measurements are then re-introduced as prior information into the representational AI component, related to hypothesis generation, so it either refines the just submitted hypothesis or encourages the exploration of new and untested ones. The uncertainty measure is responsible for informing whether refinement is needed or not. High uncertainty about the AI falsifiability prediction is interpreted as need for hypothesis refinement. This iterative process can be executed for multiple rounds, until we reach refined and non-falsifiable new knowledge.

The hypotheses that could not be falsified by many climate models may provide us new knowledge about climate-water systems. It could also, in turn, guide new investigations, collection of additional data, and adjustment of existing climate models.

The development of CHANGE will accelerate the knowledge discovery process, allowing for rapid discard of unlikely supported hypothesis about the underlying climate-water systems.

Computational resources and software: CHANGE can be integrated with existing climate analysis software frameworks and tools, such as the DOE’s Earth System Grid Federation (ESGF) and Community Data Analysis Tools (CDAT). Embedded within these systems, the AI will be able to automatically pull in data from observational platforms and climate model simulations, including the E3SM, to mine for new relationships and test new climate hypotheses.

The software will be made available to the community as an open-source tool. Climate model and observations will be stored in the cloud so researchers could easily run experiments with it. We envision the creation of an online platform in which researchers could post hypothesis generated from their own experiments. The community will then have access to a large pool of generated hypothesis.

Given the modular nature of our proposed framework, it could be reused for other climate systems aside from water systems.

Partners/Experts:

- This white paper is supported by Prof. Arindam Banerjee from University of Illinois Urbana-Champaign. Prof. Banerjee is an AI expert with long experience developing AI techniques for climate sciences. Prof. Banerjee has expressed an interest in participating in AI4ESP workshops and discussions.

References:

[Zhang and Lin, 2018] Zhang, S. and Lin, G. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217), p.20180305, 2018.

[Devlin *et al.*, 2019] Devlin, J. and Chang, M. and Lee, K. and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[Hasanzadeh *et al.*, 2020] Hasanzadeh A, Hajiramezanali E, Boluki S, Zhou M, Duffield N, Narayanan K, Qian X. Bayesian graph neural networks with adaptive connection sampling. *Proceedings of the International Conference on Machine Learning* (pp. 4094-4104), 2020.