Using machine learning and artificial intelligence to improve model-data integrated earth system model predictions of water and carbon cycle extremes

Jinyun Tang [1], William J Riley [1], Qing Zhu [1], Trevor Keenan [1, 2]
[1]Lawrence Berkeley National Laboratory
[2] University of California, Berkeley

Focal Area(s)

The research proposed here focuses on improving the predictive power of the land component of earth system models (ESMs) using (1) model-data fusion enabled by machine learning (ML) and artificial intelligence (AI), (2) predictive modeling through the combination of ML, AI, and big-data (comprising both model output and observations), and (3) insight of ESM structure and process mechanisms gleaned from complex data using ML and AI.

Science Challenge

Current efforts on water and carbon cycle benchmarking and improving ESM predictions focus on how well models capture (1) snapshots of climatology (e.g., the spatial pattern of land surface evapotranspiration), (2) time series of aggregated variables (e.g., interannual variability of net land carbon fluxes), and (3) one-vs-one variable correlations (e.g., the relationship between precipitation and evapotranspiration), all of which are less than three dimensional. However, ESM predictions are by nature of high dimension, beyond those spanned by space and time, when variables like vegetation diversity and human water use are considered. Moreover, in 10 years, with improved spatial-temporal resolution and the inclusion of more mechanistic processes, ESMs will very likely output more diverse data streams at much larger volume. Meanwhile, thanks to technological advancements, the amount of observations will also increase dramatically, in both type and spatial-temporal coverage. Current benchmark and model-data integration paradigms and methods are insufficient to address this big-data challenge. Further, current approaches are not of sufficient specificity or fidelity for evaluation at fine spatial resolutions (e.g., 1 km), nor do they provide comprehensive understanding of the casualty relationships that affect climate extremes. To address these challenges, research is proposed here to (1) make better uses of multiple scales of observations to concurrently analyze, evaluate, and reduce ESM uncertainty, and generate process knowledge of terrestrial processes, (2) achieve the ability to clearly integrate diverse observations, ML and AI, theory, and model predictive capabilities, (3) obtain robust quantification of multivariate functional relationships (e.g., net primary productivity to precipitation, temperature, and radiation) under a wide range of environmental conditions, and (4) provide high fidelity prediction and understanding of climate extreme events at fine spatial resolutions.

Rationale

ESM predictions are uncertain because (1) the earth system comprises many components, including atmosphere, land, ocean, biosphere, cryosphere, human activities, etc., each of which is insufficiently monitored and understood; (2) when the insufficient understanding of these earth system components are combined with the limited spatial-temporal resolution of ESMs,

DOI: https://doi.org/10.2172/1769794
DOE BER, AI4ESP 2021 #AI4ESP1130 (https://www.ai4esp.org)

parameterizations of many subscale processes are prerequisite, but such parameterizations are highly uncertain; and (3) existing numerical infrastructure for model-data integration is not able to use observations to robustly and effectively constrain the many process parameterizations.

In 10 years, with the advancement of computing power and more comprehensive model representation (e.g., cloud-resolving atmospheric models, ecosystem demography models, eddy-resolving ocean models, etc.), ESM outputs are expected to increase exponentially in volume size and complexity. Accompanying this, observations are also expected to increase rapidly. Therefore, current methods for integrating observations and model output, and quantifying model predictability of climate and ecosystem patterns and events need to be upgraded in both runtime efficiency and capability of blending diverse data streams.

Taking advantage of the strengths of ML and AI techniques in efficient pattern (including causality) detection, blending variable data categories, and uncertainty quantification, ESMs will be improved significantly in numerical efficiency, uncertainty reduction, and impact assessment at high spatial-temporal resolutions. In addition, because earth system modeling has unique characteristics (such as strong constraints from first principles such as mass and energy conservation) that are not usually found in cognitive computing where most recent ML and AI techniques were developed and applied, addressing the new challenges in earth system modeling will further boost the power of ML and AI techniques and help their applications in other fields.

Narrative

We propose here that ML and AI techniques can be leveraged to extend traditional ESM benchmarking to include high-dimensional dynamic data-benchmark, model structural analysis, uncertainty reduction, generation of process knowledge, and state-of-the-art re-analysis products. Such an approach would more clearly and comprehensively allow the integration of observations, machine learning, first principle-based theory, and model predictive capabilities, with a focus for benchmarking the terrestrial water and carbon cycle under extreme conditions. Moreover, by careful version control, documentation, and storage of the research tools and data through existing DOE investments in ESS-Dive, Earth System Grid Federation, AmerifFlux etc., and similar investments by other agencies, we can use the FAIR (Findable, Accessible, Interoperable, Reusable) principles the ensure the research results are always reproducible, and their gradual improvement can be chronically tracked.

High-dimensional dynamic data-benchmark will be essential for evaluating the transient performance of ecosystem models (e.g., the FATES model integrated within E3SM) that can simulate plants with as many as hundreds of cohorts. ML and AI can help to make full use of the hyperspectral remote sensing data to constrain modeling hypotheses about trait-based parameterization of competitive and symbiotic relationships among diverse plants, and even microbes. The same approach can be applied to conduct event-based benchmarks of droughts and floods, enabling better understanding of their life cycles and climatological distributions. In particular, it may provide opportunities to combine benchmarks of atmospheric dynamics, ecosystems dynamics, and oceanic dynamics to gain a higher than 4-dimensional understanding of model predictability and bias that cannot be achieved by focusing on individual components.

Regarding analysis applications under extreme events, ML approaches (e.g., transfer entropy quantification, reservoir causality detection, and Convergent cross mapping) can be used to quantify causal-effect relationships between and among model components (e.g., Yuan et al. 2021, Huang et al. 2020). The modeled causal process network can be used to identify where structural problems exist (with respect to observational process networks) or further measurements could help reduce uncertainty. When similar approaches are applied to observations and model predictions, similarities in system organization, process importance, and uncertainty quantification can be directly applied as benchmarks. For example, Liu et al. (2019) used this approach to evaluate CMIP models for their fidelity to observed interactions between land-surface conditions, sea surface temperatures, and precipitation.

For uncertainty reduction, ML approaches can be used to reduce prediction uncertainty from climate forcing by merging simulations with observations (e.g., apply Bayesian model averaging as has been done in weather forecasting, or emulate process-based models and merge them with observed climate forcings). Uncertainty in a simulation ensemble with a wide range of model structures can be reduced by surrogating physical models with ML/AI and then conducting extensive training and performance evaluation using observational data to identify and correct the model biases.

On using ML to improve process understanding, integrating ML approaches for automated model emulation would allow for deeper integration of models and observations than is currently possible. In particular, a much broader phase space of model forcing could be evaluated, parameter estimation and uncertainty could be generated, and the process network under wide background climate conditions could be analyzed. Specific model components could be targeted, and links between them could be mapped using causality-enabled ML/AI models. Such an approach could lead to a meta-model, where instead of only analyzing ensemble mean predictions, the machine learning results could be used to evaluate the models' performances and bias-correct predictions based on that evaluation. If this approach were taken for model components across the ensemble, specific process representations that are most strongly supported by observations could be assembled to form an optimized ensemble-ML model. Over time, as individual process representations were improved, new emulators of those processes could be developed and integrated into the ensemble-ML model. Analyses of the coupled system responses, and comparisons with new observations, could proceed in a much more integrated manner than is currently possible. Such an approach would be particularly useful for extreme event prediction, where mechanistic representations are often insufficient.

Finally, by aid of ML and AI techniques, we will be able to merge ensemble model simulations with diverse measurements across observational networks and across different spatial and temporal scales to produce new predictions and re-analysis products. This approach will generalize the emergent constraint approach that is currently used in model benchmarking into a hyperspace formed both by data and models (e.g., FLUXNET-MTE).

References

Huang, Y., Z. T. Fu, and C. L. E. Franzke (2020), Detecting causality from time series in a machine learning framework, *Chaos*, *30*(6), doi:Artn 063116 10.1063/5.0007670.

Yuan et al. (2021), Deforestation reshapes land-surface energy-flux partitioning, *Environ. Res. Lett.* 16 024014.

Liu, B. Y., Q. Zhu, W. J. Riley, L. Zhao, H. X. Ma, M. Van Gordon, and L. Larsen (2019), Using Information Theory to Evaluate Directional Precipitation Interactions Over the West Sahel Region in Observations and Models, *J Geophys Res-Atmos*, *124*(3), 1463-1473, doi:10.1029/2018jd029160.