White Paper to Advance an Integrative Artificial Intelligence Framework for Earth System Predictability: AI4ESP

Title: Machine-Learning-Assisted Hybrid Earth System Modelling

Authors/Affiliations
Istvan Szunyogh
Professor
Texas A&M University
Department of Atmospheric Sciences

Edward Ott
Distinguished University Professor
University of Maryland
Department of Physics and Department of Electrical Engineering

Brian Hunt
Professor Emeritus
University of Maryland
Department of Mathematics

Focal Area: We propose a novel hybrid modeling approach that combines a physics-based, numerical model of the Earth system with a machine learning model. We also envision to integrate the hybrid model training with a data assimilation cycle, so that new observations update both the state estimate used as initial conditions and the machine learning model parameters.

Science Challenge: The proposed approach is envisioned to addresses the following challenges of Earth system modeling: (1) providing a scalable, highly efficient computational approach to combine numerical modeling and machine learning, (2) integrating "online" ML training into the data assimilation cycle (3) significantly improving the simulation and prediction of Earth system processes that are heavily influenced by the parameterization schemes of the numerical model, which include processes that involve or directly affect the water cycle.

Narrative

We propose the development of new methodology for the augmentation of a physics-based prediction system (community Earth system model) with a *machine learning (ML)* component. We envision the approach to improve the simulation and prediction of the Earth system through better utilization of the available observations. The key feature of the methodology is that all forecasts, including the short-term forecasts that define the background state estimate in data assimilation, are hybrid forecasts interactively combining forecasts by the original physics-based, numerical model and the ML model. In the hybrid-model based forecast system, analyses are used, not only to define the initial conditions of the forecasts, but also to correct the forecasts for the effects of model errors. Such corrections are expected to improve, in addition to the

DOI: https://doi.org/10.2172/1769745
DOE BER, AI4ESP 2021 #AI4ESP1129 (https://www.ai4esp.org)

accuracy of the forecasts, the quality of the analyses by improving the quality of the background. ***Our proposed hybrid methodology will not require making changes to existing numerical models and data assimilation systems*****,** and the resulting hybrid model will be adaptable to future, improved versions of the numerical model by retraining the ML component.

The term *hybrid model* usually refers to a physics-based, numerical model, in which the parameterization schemes are obtained by ML (e.g., Chevallier et al. 1998; Krasnopolsky et al. 2005, Rasp et al. 2018; Brenowitz and Bretherton 2018 and 2019; Yuval and O'Gorman 2020). We **propose a different approach**, which was introduced by Pathak et al. (2018) and further developed by Wikner et al. (2020) for large, complex, spatiotemporal physical systems. In this approach, the physics-based numerical model is used as a black-box to provide a short-term prediction for each 'time step' of the hybrid model. (The hybrid model time step is typically longer than the numerical integration time step of the physics-based model.) The hybrid-model state at the end of the time step is obtained by an optimal combination of the numerical model prediction and a *reservoir computing (RC)* (e.g., Lukosevicius and Jaeger 2009; Lukosevicius 2012) based ML model prediction. Model forecasts are obtained by the iterative application of the 'one-time-step' hybrid model. The parameters of the 'optimal combination' are determined by training the hybrid model on a long time series of archived analyses of past states of the system. Formally, the training is done by minimizing a quadratic cost-function that measures the difference between the one-time-step hybrid model predictions and the training data (analyses) over the training period. An important property of the proposed approach is that it is highly scalable for a massively parallel computer architecture, which allows for a highly efficient training of the ML component for a high dimensional system such as a state-of-the-art Earth system model. We envision to further develop the hybrid approach by integrating "online" ML training into the data assimilation cycle, so that new observations update both the state estimate used as initial conditions and the ML model parameters. Preliminary work in Wikner et al. (2021), using offline training, shows the potential benefits of integrating ML training with the data assimilation cycle.

Our group has already made significant progress towards developing a prototype system based on the proposed approach for a reduced resolution atmospheric global circulation model. Results for the ML component of this prototype were presented in Arcomano et al. (2020). We showed that the ML model outperformed the numerical model in the prediction of the near-surface atmospheric state variables. The advantage of the ML model was particularly large for the state variable that described atmospheric moisture. A full hybrid model version of the prototype already exists. Preliminary results with this system, which will be described in a soon-to-be-submitted paper (Arcomano et al. 2021), show that the hybrid model vastly outperforms both the numerical and the pure ML model. Based on these highly promising early results, we are confident that the proposed approach would also improve the simulation and prediction of processes that are heavily influenced by the parameterization schemes in a state-of-the-art numerical model of the Earth system. As many of these processes involve or directly affect the water cycle, we believe that the proposed approach has a great potential to improve the simulation and prediction of the water cycle.

References

Arcomano, T., I. Szunyogh, A. Wikner, J. Pathak, B. Hunt, and E. Ott, 2021: A hybrid modeling approach to weather prediction that combines machine learning with numerical modeling (to be submitted).

Arcomano, T., I. Szunyogh, J. Pathak, A. Wikner, B. Hunt, and E. Ott, 2020: A machine learning-based global atmospheric forecast model. *Gephys. Res. Lett.,* **47,** e2020GL087776.

Brenowitz, N. D., and C. S. Bretherton, 2019: Spatially extended tests of a neural network parameterization trained by coarse-graining. *JAMES*, 2019MS001711.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.,* **45,** 6289-6298.

Chevallier, F., F. Cheruy, N. A. Scott, and A. Chedin, 1998: A neural network approach for a fast and accurate computation of a longwave radiative budget. J. Appl. Met., 37, 1385-1397.

Krasnopolsky, V. M., M. S. Fox-Rabinovitz, and D. V. Chalikov, 2005: New approach to calculation of atmospheric model physics. *Mon. Wea. Rev.,* **133,** 1370-1383.

Lukosevicius, M., 2012: A practical guide to applying echo state networks. In *Neural networks: Tricks of the trade (2nd edition),* Eds. G. Montavon, G. B. Orr, and K.-R. Müller. Springer, Heidelberg, Germany, 659-686.

Lukosevicius, M., and H. Jaeger, 2009: Reservoir computing approaches to recurrent neural network training. *Computer Science Review,* **3,** 127-149.

Wikner, A., J. Pathak, B. Hunt, I. Szunyogh, M. Girvan, and E. Ott, 2021: Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components. https://arxiv.org/submit/3606445.

Wikner, A., J. Pathak, B. R. Hunt, M. Girvan, T. Arcomano, I. Szunyogh, A. Pomerance, and E. Ott, 2020: Combining machine learning with knowledge-based modeling for scalable forecasting and subgrid-scale closure of large, complex, spatiotemporal system. *Chaos,* **30,** 053111.

Yuval, J., P. A. O'Gorman, 2020: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications,* **11,** 3295.