AI-Based Integrated Modeling and Observational Framework for Improving Seasonal to Decadal Prediction of Terrestrial Ecohydrological Extremes

Jiafu Mao[1], Yaoping Wang[2,1], Dan Ricciuto[1], Salil Mahajan[3], Forrest Hoffman[3], Xiaoying Shi[1], and Giri Prakash[1]

[1]Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA; [2]Institute for a Secure and Sustainable Environment, University of Tennessee, Knoxville, TN, USA; [3]Computational Sciences and Engineering Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Focal Areas

(1) Insight gleaned from complex data (both observed and simulated) using artificial intelligence (AI), big data analytics, and other advanced methods, including explainable AI and physics- or knowledge-guided AI

(2) Data acquisition and assimilation enabled by machine learning, AI, and advanced methods including experimental/network design/optimization, unsupervised learning (including deep learning), and hardware-related efforts involving AI (e.g., edge computing).

Science Challenge

Improving the prediction of terrestrial ecohydrological extremes (TEE) (e.g., the extremes of evapotranspiration, soil moisture, streamflow, terrestrial water storage) at seasonal to decadal (S2D) scales requires developments in the physics, parameters, and resolution of Earth system models (ESMs) and in the methods of initializing models and post-processing the predictions. Such developments involve large ensemble ESM simulations, effective diagnostic methods for sources of prediction error, and representative site selection for new observations. This white paper proposes to develop an AI-based surrogate model for the Energy Exascale Earth System Model (E3SM) to reduce the computational requirements of the ensemble S2D simulations of TEE; use the AI-based regression and feature detection methods to diagnose and remove the E3SM TEE prediction errors; and develop an AI-based procedure for the E3SM to guide the development of an observational network for TEE to reduce uncertainty and bias in E3SM TEE predictions.

Rationale

Developing an integrated framework for the TEE S2D predictions using the fully-coupled E3SM will serve many societal needs (e.g., management of flood risks to energy infrastructure) and benefit the prediction of other sub-systems of the Earth. However, identifying an optimal setup of the influential factors (e.g., initialization method, parameterization) for such a prediction system would require a large ensemble of E3SM simulations that are computationally prohibitive. AI-based surrogate models can be trained on much smaller ensembles to effectively and efficiently emulate the behavior of the original model over a limited range of outputs using the parameters and other inputs to the default model (Lu et al., 2018; Sargsyan et al., 2020). Establishing surrogate models specially for the TEE S2D predictions from E3SM would thus drastically reduce the computational cost of finding optimal prediction setups and aid subsequent applications, including the development of recalibration methods (RMs).

Even under an optimal prediction setup, the E3SM S2D prediction of TEE may contain errors that typically are corrected using various RMs. Some RMs are wholistic (e.g., linear scaling to remove

DOI: https://doi.org/10.2172/1769666
DOE BER, AI4ESP 2021 #AI4ESP1089 (https://www.ai4esp.org)

past bias), and others are based on breaking the errors into different components (e.g., initialization drift, biases in simulated oceanic oscillations or the responses to external forcings) and then correcting each error source accordingly (e.g., Smith et al., 2020). The latter approach provides better process-oriented diagnostics of prediction errors and would benefit from the use of AI methods that can identify the key covariates for each error component, thereby leading to more adaptive recalibration as well as potential information for future E3SM development. Such AI methods can be developed from the conceptual framework of existing studies on the relationships between oceanic oscillations and predictability (Zhu et al., 2020) and on the detection and attribution of natural and anthropogenic forcings (Sippel et al., 2020). But these existing studies often use linear statistical methods and do not capture nonlinear relationships as robustly as AI methods do.

Accurate and adequate site measurements of TEE are crucial for the evaluation and recalibration of the E3SM S2D prediction system. However, similar to other leading ESMs, the E3SM is still insufficiently constrained in many parts of the world (e.g., Arctic, deforested areas) because of limited availability of continuous and representative TEE observations. The AI-based improved E3SM TEE prediction and quantification of prediction errors would enable more accurate identification of the regions, timescales, and TEE components that need more observations. Moreover, AI-based methods like the representativeness analysis (Hoffman et al., 2013) can facilitate setting up cost-effective observational sites/periods that are complementary to or even outperform existing observations (e.g., the ARM).

Narrative

The following tasks are proposed to address the identified research needs and priorities:

Task 1: Conduct the E3SM-based hindcast experiments and develop surrogate model for identifying optimal prediction setup

Since current and expected simulations in the E3SM v1 and v2 campaigns do not include hindcast experiments, a series of E3SM ensemble simulations will first be performed to test various S2D hindcast configurations (Table 1) following the IPCC protocol (Boer et al., 2016). One surrogate model will be trained across all the ensemble simulations using the spatiotemporal TEE fields as the target and using the initial states of atmosphere, ocean and land surface, external forcings, and hindcast setups as the covariates. Dummy variables (Draper and Smith, 1998) will be used to convert the qualitative setups (e.g., model physics, initialization method) into numeric values; and mathematical methods for the surrogate model will be developed from those initially applied on the land component of E3SM (ELM) (Sargsyan et al., 2020). After the surrogate model is trained, the optimal E3SM setup having the highest TEE performance will be determined from numerous surrogate model runs with various combinations of hindcast setups and parameter perturbations. The optimal setup will then be reapplied onto the original E3SM to generate ensemble hindcasts (H1), again using the time range in Table 1, to verify that the prediction performance is improved. The performance (e.g., TEE timing and frequency-intensity relationship) of the surrogate model and all the E3SM simulations will be systematically evaluated using the latest TEE observations (e.g., Wang et al., 2021).

Table 1. Potential setups for E3SM hindcast experiments.

Resolution	Physics	Parameters	Initialization method	Time range
~100 km, ~25 km	W/o active ocean and ice, w/o satellite phenology scheme for ELM	Selected TEE parameters (e.g., stomatal conductance, surface/subsurface drainage, and convective parameters)	Existing reanalysis or E3SM-DART data assimilation (Zhang et al., 2020)	10-year hindcasts with 10 ensemble members, initialized every year from 1960 to present

Task 2: Develop the AI-based model and RM for quantifying the TEE prediction errors of E3SM

The remaining TEE errors in the optimized E3SM H1 from Task 1 will be conceptualized as the composite of (1) initialization drift, (2) errors in the natural climate variability modes, (3) errors in the TEE responses to natural climate variability modes, (4) errors in the TEE responses to external forcings, and (5) unexplainable random noise in the errors from parts 1–4. Part 1 of the TEE errors will be quantified by fitting a smooth, increasing function of time to the residual of the prediction error at each grid after subtracting parts 2–4 from the prediction error. To identify the covariates for part 1, it is hypothesized that part 1 would be larger if the initial state of the land surface and ocean were significantly different from the probability distribution of states that occur in E3SM under similar climate regimes. To test this hypothesis, the prediction errors and the probability distribution of states will be obtained from the H1 across different initialization years (Table 1). Systematic spatial patterns in part 1 will be investigated using the AI-based classification, or regression using local land surface conditions (e.g., vegetation, topography, soil texture) as covariates. Part 2 will be quantified by comparing the natural variability in H1-simulated sea surface temperatures (SSTs) to real world datasets. Parts 3 and 4 will be quantified by comparing the E3SM responses to SSTs and external forcings, respectively, to the responses in the real world. These responses will be estimated as spatial patterns by applying the AI-based regression (e.g., Heinze-Deml et al., 2020) between the TEE and SSTs (Part 3), and between the TEE and external forcings (Part 4), using both the H1 and observational data. Further investigation into the response mechanisms (e.g., circulation patterns, partitioning of precipitation between evapotranspiration and runoff) and spatial covariates (e.g., local land surface conditions) can be achieved using the explainable AI methods (e.g., Barnes et al., 2020). Part 5 will be modeled as Gaussian random variables with zero mean and standard deviations that increase over time. After the relationships among covariates and the five error sources are identified, testing will be done to determine whether such relationships can be used to predict the TEE errors in H1. These relationships will be trained on part of H1 and tested on the remaining part. The predicted errors can then be removed from the hindcasts by subtraction or linear scaling, which fulfills the recalibration.

Task 3: Develop AI-based method to identify cost-effective observational network for reducing the TEE prediction uncertainty

Task 3 will perform E3SM S2D forecasts from the present to 10 years into the future using the optimal setup identified in Task 1 to identify the regions, depths (for soil moisture), and metrics (e.g., timing and frequency-intensity relationship) over which TEE are expected to occur most frequently or change significantly in the future. The RM analysis performed in Task 2 will also reveal which large prediction errors in TEE occur as results of biased E3SM responses or processes. Guided by this E3SM-predicted TEE and associated error information, targeted observations can then be guided and collected to enable more accurate evaluation of the E3SM TEE performance and better constraint of relevant model processes. To achieve cost-effective design of new networks focusing on the TEE observations, Task 3 will further apply the representativeness analysis to determine specific locations that best capture the TEE metrics and prediction errors of the identified hot-spot regions and depths (Hoffman et al., 2013). The density of the observational sites and specific time periods of the observations required to achieve representativeness within pre-defined uncertainty intervals will be determined by geostatistical methods. Cost-based optimization methods that consider the site accessibility (e.g., distance to the nearest road or nearest human settlement, topographical complexity), budget constraints, and leverageable existing observations (e.g., soil moisture and LASSO from ARM, evapotranspiration from AmeriFlux and NEON) will be applied to facilitate final selection of the observational sites. It is envisioned that the workflow proposed in Task 3 can be further used to establish an adaptive data collection framework. At the beginning of each new forecast period, the added value of new sites/observational periods identified during the previous forecast can be reassessed by examining whether they capture different types of information better than do the older observations. These re-assessments and the new forecast can then be used to determine whether any addition or removal of observational locations/timings should be made during the next forecast period.

References

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., and Anderson, D. 2020. Indicator patterns of forced change learned by an artificial neural network, J. Adv. Model. Earth Syst., 12(9), https://doi.org/10.1029/2020MS002195.

Boer, G. J., et al. 2016. The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9(10), 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

Draper, N. R. and Smith, H. 1998. Dummy variables, in Applied Regression Analysis, pp. 299–326, Wiley.

Heinze-Deml, C., Sippel, S., Pendergrass, A. G., Lehner, F., and Meinshausen, N. 2020. Latent Linear Adjustment Autoencoders v1.0: A novel method for estimating and emulating dynamic precipitation at high resolution, Geosci. Model Dev. Discuss, n.a., https://doi.org/10.5194/gmd-2020-275.

Hoffman, F. M., Kumar, J., Mills, R. T., and Hargrove, W. W. 2013. Representativeness-based sampling network design for the State of Alaska, Landscape Ecol, 28(8), 1567–1586, https://doi.org/10.1007/s10980-013-9902-0.

Lu, D., Ricciuto, D., Stoyanov, M., and Gu, L. 2018. Calibration of the E3SM land model using surrogate-based global optimization, J. Adv. Model. Earth Syst., 10(6), 1337–1356, https://doi.org/10.1002/2017MS001134.

Sargsyan, K., Safta, C., and Ricciuto, D. M. 2020. Dimensionality reduction and physics-informed recurrent neural networks for climate land models, San Francisco, CA, USA, https://agu.confex.com/agu/fm20/meetingapp.cgi/Paper/762684.

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E., and Knutti, R. 2020. Climate change now detectable from any single day of weather at global scale, Nat. Clim. Chang., 10(1), 35–41, https://doi.org/10.1038/s41558-019-0666-7.

Smith, D. M., et al. 2020. North Atlantic climate far more predictable than models imply, Nature, 583, 796–800, https://doi.org/10.1038/s41586-020-2525-0.

Wang, Y., Mao, J., Hoffman, F., and Jin, M. 2021. A set of global gridded merged long-term (1971-2016) soil moisture products, To be submitted, n.a.

Zhang, S., Zhang, K., Wan, H., Anderson, J., Raede, K., and Sun, J. 2020. Ensemble hindcasts using the E3SM atmosphere model and the Data Assimilation Research Testbed, San Francisco, CA, USA, https://agu.confex.com/agu/fm20/meetingapp.cgi/Paper/700191.

Zhu, S., Chen, H., Dong, X., and Wei, J. 2020. Influence of persistence and oceanic forcing on global soil moisture predictability, Clim Dyn, 54, 3375–3385, https://doi.org/10.1007/s00382-020-05184-8.