An AI-Enabled MODEX Framework for Improving Predictability of Subsurface Water Storage across Local and Continental Scales

Dan Lu (lud1@ornl.gov)[1], Eric Pierce[1], Shih-Chieh Kao[1], Guannan Zhang[1], Juan Restrepo[1], David Womble[1], Li Li[2], and Daniella Rempe[3]

[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA; [2]Pennsylvania State University, University Park, PA, USA; [3]University of Texas at Austin, Austin, TX, USA

Focal Area

(2) Predictive modeling through the use of AI techniques and AI-derived model components. **(3)** Insight gleaned from complex data using AI, big data analytics, and other advanced methods. We propose an AI-enabled model-experiment (MODEX) framework to improve the predictability of subsurface water storage (SWS) from local to conus scales in a changing environment by taking advantage of DOE's observation and simulation capabilities, as well as to inform the model and the observation development.

Science Challenge

SWS, including the root zone storage and the rock moisture stored in weathered bedrock beneath the soil, is a significant component of the terrestrial hydrologic cycle and plays a critical role in droughts. SWS regulates the timing and magnitude of runoff and evapotranspiration fluxes; SWS dynamics influence biogeochemical cycling of carbon and nutrients; and SWS availability controls aboveground ecosystems by controlling the dominant vegetation and affects atmospheric circulation by regulating transpiration fluxes. However, because of the difficult accessibility of the underground, hydrologic properties and dynamics of SWS are poorly known. Limited direct observations of SWS exist, and accurate incorporation of SWS dynamics into Earth system land models (ELMs) remains challenging. Here, we seek to describe how an AI framework can help answer the following questions: (1) What can we learn about SWS from data (including model-simulated and real measurements)? (2) How does SWS perform as a mediator of groundwater and streamflow and as a reservoir to vegetation and thus to the atmosphere? (3) How does SWS change across local and continental scales in a changing environment? Addressing these questions will improve the predictability and understanding of SWS and therefore the integrative water cycle and associated water cycle extremes.

Rationale

Improving predictability of SWS requires a large number of data, comprehensive model representation of SWS dynamics, and sophisticated data-model integration methods for accurate prediction and effective uncertainty quantification. However, only limited direct measurements of SWS are available and current ELMs have inadequate processes representation of SWS dynamics, although an increasingly broad collection of indirect observations exist and ELMs have increasing resolution and complexity. Additionally, existing data assimilation methods are not powerful enough to incorporate diverse data for prediction and are not computationally efficient enough to integrate data streams for updating prediction, and they lack capabilities to quantify various sources of uncertainties (including meteorological forcing and geological structure uncertainties that control SWS and model process and parameter uncertainties that relate to modeling) and to identify the data and model limits to improve the prediction.

DOI: https://doi.org/10.2172/1769675
DOE BER, AI4ESP 2021 #AI4ESP1082 (https://www.ai4esp.org)

The main limitation of current data assimilation methods lies in that they focus on the model domain instead of the data domain. They use observations to first estimate model parameters and then use the calibrated models for prediction. With the increase of resolution and complexity of ELMs, these methods become computationally demanding and, at times, infeasible. Also, because the models tune parameters to fit the observations by compensating model errors, they are unable to analyze the model and data limits to predictability. In this white paper, we respond to the emerging need to better understand SWS and the challenges of SWS predictability with an AI-enabled MODEX framework. We focus on data and directly predict SWS from a variety of data, including model-simulated data, satellite data of geophysical images, and field measurements, such as streamflow, topography, permeability, porosity, groundwater table, and rooting depth from DOE-supported datasets. We leverage AI's power in data analytics and predictive analytics to link models with diverse data for prediction and to analyze model and data limits to inform the model and data development, thus improving predictability and understanding of SWS and its role in integrative water cycle and associated water cycle extremes.

Narrative

The proposed framework consists of three interconnected capabilities (Figure 1): (I) a data-model informed prediction that links model and data and sufficiently extracts their information for prediction with considering various sources of uncertainty; (II) a model-driven data collection that analyzes data limits to predictability, identifies informative data, and guides data investment to enhance predictive skill, and (III) a data-driven model improvement that analyzes model limits to predictability, identifies model deficiency, and complements missing physics with AI models to advance model development.

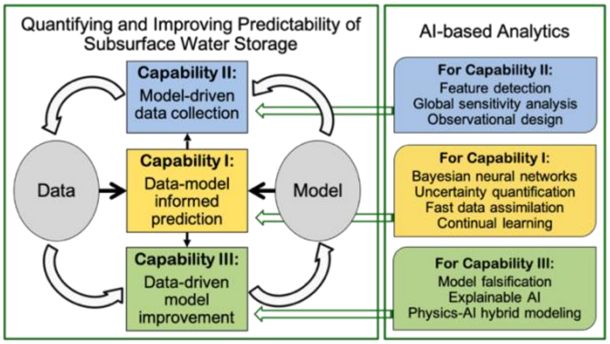
Quantifying and Improving Predictability of Subsurface Water Storage
Capability II: Model-driven data collection
Data
Capability I: Data-model informed prediction
Model
Capability III: Data-driven model improvement
AI-based Analytics
For Capability II: Feature detection Global sensitivity analysis Observational design
For Capability I: Bayesian neural networks Uncertainty quantification Fast data assimilation Continual learning
For Capability III: Model falsification Explainable AI Physics-AI hybrid modeling

Figure 1. An AI-enabled MODEX framework for advancing understanding and predictability of SWS. This novel framework allows for considering various sources of uncertainty, linking diverse data with model for prediction improvement and uncertainty reduction, and analyzing the data and model limits to inform the data and model development by leveraging AI, exascale computing and edge computing.

Capability I: A Novel Data-Model Informed Prediction

Our prediction framework focuses on leveraging AI techniques to learn a direct relationship between data variables (in which we have observations including direct observations of SWS and indirect streamflow) and prediction variables (i.e., SWS at the selected locations and times), and then deploys this learned data-prediction relationship (i.e., an AI model) and uses the actual observations for prediction. In this framework, the role of models is reconsidered in which models are forward-simulated to generate samples of data variables and prediction variables to establish their relationship instead of being inversely calibrated to match the observations in the traditional data-assimilation methods. This new formulation has several benefits. (1) It allows for considering a variety of sources of uncertainty simultaneously in the forward simulations for improving SWS predictability across a broad range of geological and climatic settings. (2) It uses observations to directly reduce prediction uncertainty based on the learned data-prediction relationship without computationally challenging parameter optimization, which enables efficient prediction and fast data assimilation. (3) It uses online training for the ensemble forward simulations and offline learning for assimilating observations [1]. This strategy can leverage the exascale computing for parallel simulations and the edge computing for continually updating predictions from the observation streams. A collection of AI techniques and analysis is proposed to implement this capability, including Bayesian deep neural networks [2] to learn the data-prediction relationship, surrogate modeling [3] to accelerate the forward simulation, dimension reduction and feature detection to extract sample information, and continual learning to assimilate data streams.

Capability II: Model-Driven Data Collection

We propose to use feature detection and sensitivity analysis to guide the spatiotemporal data acquisition. We will first use feature detection techniques to identify where SWS is likely to significantly affect hydrologic fluxes and state variables and what types of data and how much information are missing to improve the prediction. Then, we will conduct a two-way global sensitivity analysis [4] to identify key data variables and locations that can constrain those uncertain parameters and processes that have a vital impact on predictions. Finally, we will perform a value of information analysis [5] for the cost-effective observational design. These analyses will be performed in the reduced dimensions of the data and prediction variables. Our new framework makes this dimension reduction feasible and effective because for SWS prediction, the data variables and prediction variables are usually time series or spatial maps whose dimensions can be reduced without much loss of information.

Capability III: Data-Driven Model Improvement

Model falsification and explainable AI will be used to inform the SWS dynamics implementation in ELMs. We will first perform model falsification to analyze the consistency between the model generated data samples and the actual observations. If the data samples are inconsistent with the observations by showing the observations outside the sample clouds, the models are falsified, and the falsified models cannot make effective prediction in the out-of-observation regime (such as different geological and climatic settings). Then, we will use explainable AI techniques to analyze the model deficiency and detect the missing processes. Finally, we will build a data-driven AI model [6] to compensate the missing SWS dynamics in the ELMs for a closure simulation in the use of physics-AI hybrid modeling. Capability I will inform Capabilities II and III, which will advance Capability I.

Use Cases

We propose an initial exploration and demonstration of the framework for SWS predictability in four intensively studied watersheds relevant to the Earth and Environmental Systems Sciences Division with diverse geology and climate: Shale Hills (Pennsylvania), Walker Branch (Tennessee), Elder Creek (California), and East River (Colorado). Diverse data sources (e.g., streamflow, stream chemistry, topography, permeability and porosity, geophysical images, groundwater table, rooting depth, soil depth, and evapotranspiration, along with ELM simulation data) will provide inputs for AI analysis. After testing and refining the techniques on the local scale, we will extend the framework to a continental scale. We will provide a detailed data management to ensure the generated data are findable, accessible, interoperable, and reusable. Code packages of this activity will be open-sourced, tested on various hardware, and reusable for other Earth system problems.

References

[1] Lu, D., and Ricciuto, D. 2019. Learning-based inversion-free model-data integration to advance ecosystem model prediction. DOI 10.1109/ICDMW.2019.00049.

[2] Lu, D., Liu, S., and Ricciuto, D. 2019. An efficient Bayesian method for advancing the application of deep learning in earth science. DOI 10.1109/ICDMW.2019.00048.

[3] Lu, D., and Ricciuto, D. 2019. Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques, Geosci Model Dev, 12(5), 1791-1807.

[4] Lu, D., and Ricciuto, D. 2020. Efficient distance-based global sensitivity analysis for terrestrial ecosystem modeling. Proceedings of the 2020 IEEE International Conference on Data Mining Workshops.

[5] Lu, D., Ricciuto, D., and Evans, K. 2018. Efficient Bayesian data-worth analysis using a multilevel Monte Carlo method, Advances in Water Resources, 113, 223-235.

[6] Konapala, G., Kao, S., Painter, S. L., and Lu, D. 2020. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15(10).