Machine Learning-enabled Model-Data Integration for Predicting Subsurface Water Storage

Dan Lu
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
lud1@ornl.gov

Eric Pierce
Environmental Sciences Division
Oak Ridge National Laboratory
pierceem@ornl.gov

Shih-Chieh Kao
Environmental Sciences Division
Oak Ridge National Laboratory
kaos@ornl.gov

David Womble
Computer Sciences and Mathmatics Division
Oak Ridge National Laboratory
womblede@ornl.gov

Li Li
Civil and Environmental Engineering
Pennsylvania State University
lxl35@psu.edu

Daniella Rempe
Department of Geological Sciences
The University of Texas at Austin
rempe@jsg.utexas.edu

Abstract

Subsurface water storage (SWS) is a key variable of the climate system and a storage component for precipitation and radiation anomalies, inducing persistence in the climate system. It plays a critical role in climate-change projections and can mitigate the impacts of climate change on ecosystems. However, because of the difficult accessibility of the underground, hydrologic properties and dynamics of SWS are poorly known. Direct observations of SWS are limited, and accurate incorporation of SWS dynamics into Earth system land models remains challenging. We propose a machine learning-enabled model-data integration framework to improve the SWS prediction at local to conus scales in a changing climate by leveraging all the available observation and simulation resources, as well as to inform the model development and guide the observation collection. The accurate prediction will enable an optimal decision of water management and land use and improve the ecosystem's resilience to the climate change.

1 Introduction

Subsurface water storage (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 [4, 1]. 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 [18]. However, because of the difficult accessibility of the underground, hydrologic properties and dynamics of SWS are poorly known. Limited direct observations of SWS exist [11, 5], and accurate incorporation of SWS dynamics into Earth system land models (ELMs) remains challenging [6, 16]. Here, we seek to describe how machine learning (ML) 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

Tackling Climate Change with Machine Learning workshop at NeurIPS 2021

atmosphere? (3) How does SWS change across local and continental scales in a changing climate? Addressing these questions will improve the predictability and understanding of SWS and therefore its critical role in mitigating the impacts of climate change and associated water cycle extremes on the ecosystems and many societal sectors including energy production, health, forestry and agriculture.

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 [12]. 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 [9]. 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 [2, 7].

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 [3]. 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. To address these challenges, we propose ML-enabled model-data integration. In the proposed 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 a variety of database supported by many agencies including Department of Energy and U.S. Geological Survey. We leverage ML'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 enhancing predictability and understanding of SWS and its role in integrative water cycle and its significance in improving ecosystem's resilience to the climate change.

2 Method

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 ML models to advance model development.

2.1 Capability I: A Novel Data-Model Informed Prediction

The proposed prediction framework focuses on leveraging ML techniques to learn a direct relationship between data variables (in which we have observations including direct measurements of SWS and indirect streamflow) and prediction variables (i.e., SWS at the locations and times of interest), and then deploys this learned data-prediction relationship (i.e., a ML 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. 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 ML techniques and analysis is proposed to implement this capability, including Bayesian deep neural networks [19]

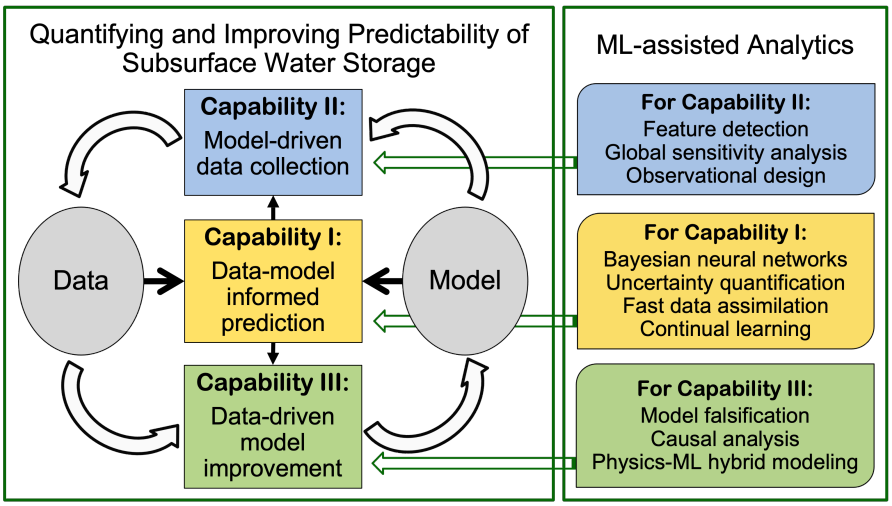

Figure 1: ML-enabled model-data integration for advancing understanding and predictability of SWS. This novel framework allows for considering various sources of uncertainty, linking diverse data with model for prediction improvement and uncertainty reduction, and analyzing the data and model limits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

to learn the data-prediction relationship, surrogate modeling [14] to accelerate the forward simulation, dimension reduction [8] and feature detection to extract sample information, and continual learning to assimilate data streams.

2.2 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 [17] 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 [20, 10] 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 greatly reduced without much loss of information.

2.3 Capability III: Data-Driven Model Improvement

Model falsification and casual analysis 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 causal analysis to explore the underlying variable interconnection from the data and generate new hypothesis [13]. Finally, we will build a data-driven ML model from the hypothesis generation to compensate the missing SWS dynamics in the ELMs for a closure simulation in the use of physics-ML hybrid modeling [15]. Capability I will inform Capabilities II and III which will in turn advance Capability I.

3 Impact and Future work

SWS is a key variable of the climate system. It constrains plant transpiration and photosynthesis, with consequent impacts on the water, energy and biogeochemical cycles. Moreover, it is a storage

component for precipitation and radiation anomalies, inducing persistence in the climate system. Finally, it is involved in a number of feedbacks at the local, regional and global scales, and plays a major role in climate-change projections. This paper proposes a ML-enabled model-data integration idea to improve the predictability of SWS. We identified four intensively studied watersheds with diverse geology and climate — Shale Hills (Pennsylvania), Walker Branch (Tennessee), Elder Creek (California), and East River (Colorado) — for demonstration of the proposed idea. 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) at these four sites will provide inputs for the ML analysis. After testing and refining the techniques on the local scale, we will extend the framework to a continental scale.

References

[1] Hao, Z., Singh, V. P., and Xia, Y. (2018). Seasonal drought prediction: Advances, challenges, and future prospects. *Reviews of Geophysics*, 56(1):108–141.

[2] Hartick, C., Furusho-Percot, C., Goergen, K., and Kollet, S. (2021). An interannual probabilistic assessment of subsurface water storage over europe using a fully coupled terrestrial model. *Water Resources Research*, 57(1):e2020WR027828. e2020WR027828 2020WR027828.

[3] Hill, M. C. (2000). *Methods and Guidelines for Effective Model Calibration*, pages 1–10.

[4] Hirschi, M., Seneviratne, S., and Alexandrov, V. (2011). Observational evidence for soil-moisture impact on hot extremes in southeastern europe. *Nature Geosci.*, 4:17–21.

[5] Ireson, A., Wheater, H., Butler, A., Mathias, S., Finch, J., and Cooper, J. (2006). Hydrological processes in the chalk unsaturated zone – insights from an intensive field monitoring programme. *Journal of Hydrology*, 330(1):29–43. Hydro-ecological functioning of the Pang and Lambourn catchments, UK.

[6] Koirala, S., Jung, M., Reichstein, M., de Graaf, I. E. M., Camps-Valls, G., Ichii, K., Papale, D., Ráduly, B., Schwalm, C. R., Tramontana, G., and Carvalhais, N. (2017). Global distribution of groundwater-vegetation spatial covariation. *Geophysical Research Letters*, 44(9):4134–4142.

[7] Li, P., Zha, Y., Shi, L., Tso, C.-H. M., Zhang, Y., and Zeng, W. (2020). Comparison of the use of a physical-based model with data assimilation and machine learning methods for simulating soil water dynamics. *Journal of Hydrology*, 584:124692.

[8] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

[9] Maxwell, R. M., Chow, F. K., and Kollet, S. J. (2007). The groundwater–land-surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. *Advances in Water Resources*, 30(12):2447–2466.

[10] Neuman, S. P., Xue, L., Ye, M., and Lu, D. (2012). Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources*, 36:75–85. Special Issue on Uncertainty Quantification and Risk Assessment.

[11] Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S., Hagimoto, Y., Kerr, Y. H., Larson, K. M., Njoku, E. G., Small, E. E., and Zreda, M. (2013). State of the art in large-scale soil moisture monitoring. *Soil Science Society of America Journal*, 77(6):1888–1919.

[12] Paniconi, C., Troch, P. A., van Loon, E. E., and Hilberts, A. G. J. (2003). Hillslope-storage boussinesq model for subsurface flow and variable source areas along complex hillslopes: 2. intercomparison with a three-dimensional richards equation model. *Water Resources Research*, 39(11).

[13] Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*, 6(1).

[14] Razavi, S., Tolson, B. A., and Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, 48(7).

[15] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204.

[16] Sahoo, S., Sahoo, B., and Panda, S. N. (2018). Hillslope-storage boussinesq model for simulating subsurface water storage dynamics in scantily-gauged catchments. *Advances in Water Resources*, 121:219–234.

[17] Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, 22(3):579–590.

[18] Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3):125–161.

[19] Wang, H. and Yeung, D.-Y. (2016). Towards bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408.

[20] Wang, Y., Shi, L., Lin, L., Holzman, M., Carmona, F., and Zhang, Q. (2020). A robust data-worth analysis framework for soil moisture flow by hybridizing sequential data assimilation and machine learning. *Vadose Zone Journal*, 19(1):e20026.