Toward Hybrid Physics - Machine Learning to improve Land Surface Model predictions

1. **Title**: Toward Hybrid Physics-Machine Learning to improve Land Surface Model predictions
2. **Authors/Affiliations**:

Geza, M., Assistant professor, Hydrology, SD Mines
Tesfa T., Scientist, Hydrology, Pacific Northwest National Laboratory (PNNL)
Liangping Li, Assistant professor, Hydrogeology, SD Mines
Qiao, M. Associate Professor, Data-Scientist/Machine Learning, SD Mines

3. **Focal Area(s):** Improving land surface model predictions using physics informed machine learning.
4. **Science Challenge**: Improving the E3SM Land Model (ELM) predictions using integrated Physics Informed Machine Learning (PIML) modeling.
5. **Rationale**:

A critical challenge for Land Surface Models (LSMs) is to simulate processes at the surface and the subsurface and their feedbacks to the atmosphere. Even using the same climate forcings, different LSMs predict different surface fluxes and soil moisture conditions due to differences in the formulations of individual processes, parameterizations, and representation of spatial heterogeneity. Ultimately, these differences contribute to the LSM prediction errors and uncertainty. This research seeks to address this challenge by coupling physics-based modeling with state-of-the-art machine learning (ML) techniques to describe complex physical and biogeochemical processes and narrow the gap between model predictions and observations.

**6. Narrative**

Physics based land surface models (LSMs) have been developed and improved over time to predict how terrestrial fluxes of carbon, water, and energy to inform other components of Earth System Models (ESMs) (atmosphere, ocean etc.). The E3SM Land Model (ELM) is the land component of the Energy Exascale Earth System Model (E3SM) that is being developed by the Department of Energy (DOE) to investigate DOE's mission-relevant science questions (Golaz et al., 2019). ELM was branched from the land model component of the Community Earth System Model (CESM1), the Community Land Model (CLM), but it has evolved significantly. Over the last decade, CLM has been developed and expanded to address weather and climate impacts on water availability, crop yields, wildfire risk, heat stress, and other ecosystem services (Bonan and Doney, 2018, Lawrence et al., 2011, Oleson et al., 2013) and has been used in various regional and global modeling efforts (Lawrence, et al., 2019). CLM predictions have been compared to satellite-derived measurements or observationally derived data syntheses, which has revealed areas and processes where models perform relatively well, and/or places where improvements could be made (Luo et al., 2012, Kelly et al., 2013, Collier et al., 2018). Branched from CLM, ELM also has been improved significantly by the DOE with new modeling capabilities such as irrigation (Zhou et al., 2020), reservoir (Voisin et al., 2017) and inundation (Luo et al., 2017; Mao et al., 2019).

Like most physically based models (PBMs), LSMs have various drawbacks that limit their use. Such models have a number of parameters, which are difficult to measure, often determined by fitting model outputs to available data through a calibration process that requires extensive effort (Arnold et al. 2012; Shen et al., 2012) and even after calibration, there are uncertainty in the model outputs (Geza, et al., 2009). Some of the limitations include a problem of equifinality or non-uniqueness of model parameter values (Beven and Freer, 2001), large number of parameters and hence the risk of over-parameterization, and the lack of complete understanding of the underlying physical and biogeochemical processes. The uncertainties in model outputs may arise from model structure associated with model itself, simplified representation of the processes, and from parameter estimates (Kavetski et al., 2002; Kavetski et al., 2006a; Kuczera et al., 2006).

DOI: https://doi.org/10.2172/1769785
DOE BER, AI4ESP 2021 #AI4ESP1049 (https://www.ai4esp.org)

An alternative approach to PBMs with fast growing application is Machine Learning (ML). In contrast to PBMs, ML-based methods recognize patterns hidden in observed data, and provide quick and direct correlation between predictors and hydrological responses without explicit descriptions of the underlying processes (Adnan et al., 2019). Several studies have shown that ML models can outperform PBMs with respect to some of the outputs (Kratzert et al., 2019). However, they may produce outcomes inconsistent with physical laws if they are not constrained. LSMs need to embrace data driven ML modeling as a new paradigm by leveraging emerging datasets to improve prediction accuracy. We propose to develop a hybrid ELM-ML model that combines the strength of physics based model and ML model. We will incorporate our understanding of physical processes to constrain ML models and reduce the need for large amounts of data compared to traditional data-driven approaches. We will make use of new ML modeling approaches to improve specific processes that are not sufficiently described using physics based modeling.

Given that neither an ML-only nor currently used LSM approach can be considered sufficient for complex scientific and engineering applications, we propose to explore the continuum between mechanistic and ML models, where both scientific knowledge and data are integrated in a synergistic manner. With expanding observational data in hydrology and water quality, merging principles from ML and physics can play an invaluable role in addressing environmental and water resources management problems. ML models, given enough data, can find structure and patterns in problems where complexity limits explicit description using PBMs. Given this ability to extract complex relationships from data, ML models appear promising for scientific problems with physical processes that are not fully understood. We hypothesize that prediction accuracy of LSMs can be enhanced by coupling physically based land surface models with a state-of the art enhanced ML models and leveraging their complementary strengths. The potential for hybrid models to describe complex physical and biogeochemical processes has not been explored in detail yet. The research aligns with Earth and Environmental Systems Sciences Division Strategic Plan particularly with integrated water cycle and data-model integration scientific grand challenges.

**Objective:** We hypothesize that coupled hybrid models can help capture the dynamics of hydrology and biogeochemistry and provide better prediction accuracy and generalizability with a smaller number of observations. Thus, the objective of the study is to combine elements of physics-based modeling with state-of-the-art data driven ML models to leverage their complementary strengths. The research will advance our understanding and modeling capabilities of mass and energy exchange at the land-atmosphere interface.

**Approach:** The research involves (1) Data acquisition for CLM and ML modeling, (2) CLM model development and calibration, (3) ML model development. Both single learners (e.g. Support Vector Machine (SVM), Deep Neural Network (DNN)) and Stacking ensemble ML approaches (e.g. Adaboost and Random Forest) will be adopted to build predictive models for evaluation. (4) Dimensionality reduction. ML-based feature selection and reduction (e.g. SVM feature elimination and deep generative models) will be employed to reduce the overhead and improve the performance. (5) Development a hybrid physics-data driven ML. Best performing ML model will be used in the hybrid structure. Novel methods will be used to combine ELM with state-of-the-art ML techniques including physics–guided loss function approach to constrain ML model with physical principles and a hybrid physics-ML model approach that involves aggregation of predictions from physics based and ML models into a single "best-estimate" forecast that minimizes bias between predictions and observations, and (6) comparing existing physics based, pure ML, and hybrid models. Comparison among the models will reveal the limitation and strength of existing and new approaches in describing surface fluxes and soil moisture conditions.

**Test site:** We will use coniferous forest site in the Pacific Northwest region of the U.S. (Wind River AmeriFlux site) to build our models. Wind River is part of the AmeriFlux eddy covariance network (Baldocchi et al., 2001), with a long record of meteorological, biological, surface flux (energy and carbon), and carbon isotope measurements for model assessment (1998–present). The site is located in the Pacific Northwest region of the United States, in the state of Washington.

**Data type and source:** Data for ML and ELM modeling such as air temperature, relative humidity, wind speed, radiation, atmospheric pressure, and precipitation can be obtained from the Wind River site. AmeriFlux repository has data on sensible heat (H), latent heat (LE), and carbon, including GPP and ecosystem respiration (ER). The data will be used for ELM model development, calibration, and validation and for training and testing ML and hybrid models.

**ELM model set up and calibration:** The ELM will be built at site level. The model will be configured to run both physics and carbon–nitrogen biogeochemistry. To improve model predictions, model parameters will be calibrated. Parameter adjustments will be based on measurements at the site. Parameters controlling various aspects of model outputs including energy fluxes, potential gross primary production, and soil moisture will be calibrated. The groundwater and water table fluctuation influences soil moisture, surface energy, and evapotranspiration (Yeh and Eltahir, 2005). Water table depth in CLM is based on simple groundwater model (Niu et. al, 2007). Other physically based analytical solutions for water table fluctuation will be evaluated and coupled with ML model to improve water table predictions. The modified groundwater module will be evaluated against the Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage change data or water table data from USGS.

**ML model set up training and testing:** Multiple single ML methods including SVM and DNN will be used to build predictive models using selected features. SVM and DNN have gained popularity in many Artificial Neural Network (ANN) dominated fields (Ahmed, et al., 2019). Ensemble stacking learning have also been demonstrated to improve the generalizability and prediction accuracy of machine learning (ML) models by combining the merits of several models. The ML models will handle supervised learning for massive data as well as semi-supervised learning tasks where data is insufficient. We will build our ML algorithms using the Python ML classification and regression training package.

**Coupling physics based model with ML model:** The proposed coupling requires passing information between the CLM and the ML model. Altering inputs, data exchange between the models, and repeatedly executing each model requires an additional ‘wrapper software platform’. Open Modeling Interface (OpenMI) (Moore and Tindall 2005; Gregersen, et al. 2007) with a wrapper software has been used to enable communication between models via data exchange and execution. This requires each model to be OpenMI compliant; various models are in the process of becoming OpenMI compliant. Thus, dynamically linking ELM to an ML model may require developing new wrapper software or using OpenMI. The ML algorithms will be built using the Python ML classification and regression training package. Use of a Python platform will facilitate full integration of ML model with the ELM.

**Model comparison:** The three models will be evaluated against observations. For ELM, we will compare the models using calibration and validation steps. For ML and the hybrid model, model performance will be evaluated during training and testing phases. Validation of ELM model or testing of ML model involves evaluating calibrated or trained models outside of calibration or training dataset using the same parameter values estimated during calibration or training.

## References

Ahmed, A.N., Othman, F.B., Afan, H.A., Elsha, A., 2019. Machine learning methods for better water quality prediction. J. Hydrol. 578, 1–18.

Adnan, R.M., Liang, Z., El-Shafie, A., Zounemat-Kermani, M., Kisi, O., 2019a. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577.

Arnold, J. G.; Moriasi, D. N.; Gassman, P. W.; Abbaspour, K. C.; and White, M. J. 2012. SWAT: Model use, calibration, and validation. Transactions of the ASSBE 55(4):1491–1508.

Baldocchi, D.D. et al., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society, 82(11): 2415-2434.

Beven, K. and Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of hydrology*, *249*(1-4), pp.11-29.

Bonan, G.B. and Doney, S.C., 2018. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. Science, 359(6375).

Collier, N.; Hoffman, F.M.; Lawrence, D.M.; Keppel-Aleks, G.; Koven, C.D.; Riley, W.J.; Mu, M.; Randerson, J.T. The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. J. Adv. Model. Earth Syst. 2018, 10, 2731–2754

Geza, M., Poeter, E.P. and McCray, J.E., 2009. Quantifying predictive uncertainty for a mountain-watershed model. *Journal of hydrology*, *376*(1-2), pp.170-181.

Golaz J., P.M. Caldwell, L. Van Roekel, M.R. Petersen, Q. Tang, J. Wolfe, and G.W. Abeshu, et al. 2019. "The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution." Journal of Advances in Modeling Earth Systems 11, no. 7:2089-2129. PNNL-SA-141816.

Gregersen, J.B., Gijsbers, P.J.A. and Westen, S.J.P., 2007. OpenMI: Open modelling interface. *Journal of hydroinformatics*, *9*(3), pp.175-191.

Kavetski, D., S. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modelling in calibration of watershed models, in Water Sci. Appl. Ser., vol. 6, edited by Q. Y. Duan, et al., pp. 49–68, AGU, Washington, D.C.

Kavetski, D., Kuczera, G. and Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water resources research*, *42*(3).

Kelley, D.I.; Prentice, I.C.; Harrison, S.P.; Wang, H.; Simard, M.; Fisher, J.B.; Willis, K.O. A comprehensive benchmarking system for evaluating global vegetation models. Biogeosciences 2013, 10, 3313–3340. Lawrence, D.M., Oleson, K.W., Flanner, M.G., Thornton, P.E., Swenson, S.C., Lawrence, P.J., Zeng, X., Yang, Z.L., Levis, S., Sakaguchi, K. and Bonan, G.B., 2011. Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems*, *3*(1).

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. and Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), pp.5089-5110.

Kuczera, G., Kavetski, D., Franks, S. and Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1-2), pp.161-177.

Lawrence, D.M., Fisher, R.A., Koven, C.D., Oleson, K.W., Swenson, S.C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D. and Kluzek, E., 2019. The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), pp.4245-4287.

Luo, X., Li, H.-Y., Leung, L. R., Tesfa, T. K., Getirana, A., Papa, F., and Hess, L. L.: Modeling surface water dynamics in the Amazon Basin using MOSART-Inundation v1.0: impacts of geomorphological parameters and river flow representation, Geosci. Model Dev., 10, 1233–1259, Mao Y., T. Zhou, L. Leung, T.K. Tesfa, H. Li, K. Wang, and Z. Tan, et al. 2019. "Flood Inundation Generation Mechanisms and Their Changes in 1953-2004 in Global Major River Basins." Journal of Geophysical Research: Atmospheres 124, no. 22:11672-11692. PNNL-SA-147906.

Moore, R.V. and Tindall, C.I., 2005. An overview of the open modelling interface and environment (the OpenMI). *Environmental Science & Policy*, *8*(3), pp.279-286.

Niu, G.-Y., Yang, Z.-L., Dickinson, R.E., Gulden, L.E., and Su, H. 2007. Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate

Experiment data. J. Geophys. Res. 112:D07103.

Oleson, K.W., Lawrence, D.M., Gordon, B., Flanner, M.G., Kluzek, E., Peter, J., Levis, S., Swenson, S.C., Thornton, E., Feddema, J. and Heald, C.L., 2010. Technical description of version 4.0 of the Community Land Model (CLM).

Shen, Z.Y., Chen, L. and Chen, T., 2012. Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: a case study of SWAT model applied to Three Gorges Reservoir Region, China. *Hydrology and Earth System Sciences*, *16*(1), pp.121-132.

Voisin, N., Hejazi, M.I., Leung, L.R., Liu, L., Huang, M., Li, H.Y. and Tesfa, T., 2017. Effects of spatially distributed sectoral water management on the redistribution of water resources in an integrated water model. *Water Resources Research*, *53*(5), pp.4253-4270.

Yeh, P. J. F., and E. A. B. Eltahir (2005), Representation of water table dynamics in a land surface scheme, Part I: Model development, J. Clim., 18(12), 1861-1880.

Zhou T., L. Leung, G. Leng, N. Voisin, H. Li, A. Craig, and T.K. Tesfa, et al. 2020. "Global irrigation characteristics and effects simulated by fully coupled land surface, river, and water management models in E3SM." Journal of Advances in Modeling Earth Systems 12, no. 10: Article No. e2020MS002069. PNNL-SA-147925.