Building Intelligent Cyberinfrastructure to Learn Iteratively from both Observations and Models for Understanding Watershed Dynamics

Authors/Affiliations

Xingyuan Chen[1], Umakant Mishra[2], Joshua B. Fisher[3], Peishi Jiang[1], Maruti K. Mudunuru[1], Alexander Sun[4], Pin Shuai[1], Sagar Gautam[2], David Moulton[5]

[1] Pacific Northwest National Laboratory, Richland, WA.

[2] Computational Biology & Biophysics, Sandia National Laboratories, Livermore, CA.

[3] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA.

[4] Bureau of Economic Geology, University of Texas, Austin, TX.

[5] Los Alamos National Laboratory, Los Alamos, NM.

Focal Area(s)

- Predictive modeling through the use of AI techniques and AI-derived model components; the use of AI and other tools to design a prediction system comprising of a hierarchy of models (e.g., AI-driven model/component/parameterization selection)

Science Challenge

Watershed processes, such as the fate and transport of sediment, carbon and nutrients across landscapes and their fluxes to water bodies (e.g., streams, rivers and lakes), have important implications for global and regional carbon and nutrient dynamics, biogeochemical functioning of terrestrial ecosystems, and soil functions. The magnitude of lateral surface/subsurface transport and fluxes of sediment, carbon and nutrients are key factors controlling the vulnerability of watersheds to climate extremes such as droughts, wildfires, and floods. Recent field observations and other scientific evidence suggest that the magnitudes of lateral transport and fluxes of sediment, carbon and nutrients are governed primarily by the spatial and vertical heterogeneity of landscape and soil properties and by pedogenic processes. However, the current generation of land surface and watershed models do not mechanistically couple the terrestrial and hydrologic systems, nor do they represent sufficiently the spatial and vertical heterogeneity of land surface and subsurface properties. On the other hand, increasing complexity of coupled watershed and land surface models requires more data to parameterize, calibrate and validate. Remote sensing (RS) provides a means to acquire spatial data and characterize their heterogeneity at the watershed scale, overcoming a major limitation associated with conventional point measurements. To improve the representation of land-surface and surface/subsurface process coupling and sub-grid heterogeneity in watershed models, it is essential to build our predictive understanding by learning from both the multi-scale multi-process modeling and diverse multi-scale data while leveraging powerful artificial intelligence (AI) techniques.

DOI: https://doi.org/10.2172/1769684

Rationale

The iteration between model and experiments (ModEx) is essential to improving the predictability of watershed models under both baseline and perturbed conditions[1]. The increase in model complexity and data volume has led to substantial increase in computational cost and exponential increase in data dimensionality that have both hampered ModEx. Meanwhile, the observational data on extremes are scarce due to their rare occurrence and practical challenges in collecting data under those conditions. Targeted data collection assisted by modeling can increase the information content of data for model improvement. An important question to address in watershed science is **what process representations and levels of mechanistic details must be captured in watershed biogeochemical models for robust prediction of the temporal and spatial patterns of solute fluxes (e.g., DOC and nitrate) before, during, and after extreme events?** Answering this question requires a systematic way to integrate data and model with varying complexity and evaluate the model performance against each other or observation data to identify gaps.

Machine learning (ML) methods have been developed to assist various components of model-data integration across scales, such as upscaling and downscaling, building surrogate models to reduce computational cost, estimating parameters through inverse modeling, and transferring mechanistic understanding across scales and domains. We are facing several challenges to boost the adoption of AI/ML methods[2]: (1) incorporating physics in ML models; (2) improving the interpretability of ML models; (3) enabling reliable extrapolations beyond the training conditions; **(4) quantifying and propagating uncertainty in model results; (5) developing publicly available benchmark training data sets that can be used to aid and test new ML methods; and (6) building a community computational platform to allow the share of ML-assisted ModEx pipelines, with easy access to pre-trained ML models** (e.g., similar to Model Zoo, https://modelzoo.co/), **standardized application ready datasets, interoperable process-based models, and supercomputing and/or cloud computing resources.** Extensive research to address the first three challenges is currently being pursued. Here, we call for significant investment to support community efforts that address the last three challenges.

Narrative

We propose to build community cyberinfrastructure to accelerate the systematic integration of multi-scale modeling with highly heterogeneous data. Deep reinforcement learning[3] (RL) techniques can be used as the overall framework of an automated, intelligent model-data integration system for ModEx. RL trains ML models to make a sequence of decisions in an uncertain and complex environment, with example well-known successes in autonomous cars and AlphaGo. ModEx shares similar objectives with RL in continuously learning from both data and models under their uncertainty until the model achieves desired predictability. Within the RL framework, models are rewarded based on the quality of their predictions and continue to improve until a reward threshold is reached. The quality of model predictions can be assessed using metrics designed for spatio-temporal system behaviors, e.g., those implemented in the International Land Model Benchmarking (ILAMB) software (www.ilamb.org). All ModEx elements, including data assimilation, inverse modeling, sensitivity analyses, and model-

informed experimental design, can be naturally designed as a sequence of ModEx decisions. Bayesian/probabilistic inference should be integrated to enable explicit representation and propagation of uncertainties across the hierarchy of models, which can inform policy optimization in RL. Graph neural networks[4] (GNN) (e.g., HydroNets[5] and Mesh R-CNN[6]) can be explored as a scaling tool for propagating information across river networks. GNNs add flexibility to deal with irregular (vs gridded) data and model outputs using unstructured meshes, which are increasingly being used to capture hot spots and hot moments in hydrodynamic and biogeochemical processes.

Generating public benchmark training data sets (similar to ImageNet, http://www.image-net.org/) that researchers can use to build better ML models is the key to advancing applications of ML in Earth science domains[2,7]. There is a unique opportunity to enhance the use of the new generation of RS products that capture components of the water cycle (precipitation, snow, soil moisture, evapotranspiration, groundwater, and runoff), as well as coupled carbon and nutrient cycle components, with increasing spatial and temporal resolutions. Training data may also be generated from process-based models. Leveraging open-source resources from federal agencies is necessary for the success of such extensive and expensive effort. For example, NASA's Earth Sciences Data Systems (ESDS) has generated high-quality training data sets that are open and easily accessible. NOAA, USGS, and other federal agencies have been maintaining extensive observation networks and are developing a large number of integrated Earth system models. Standardized data management practices would significantly increase the data usability.

Lastly, we need computational infrastructure to address longstanding challenges of complexity and heterogeneity in watershed models that would otherwise be overwhelmed by the tremendous complexity in managing software, hardware, workflows and computational cost. Addressing these challenges requires developing and maintaining open-source scientific software and ML frameworks for deploying Earth science ML and process-based models. To achieve this goal, existing frameworks can be expanded or integrated through collaborative efforts for efficiency. The Department of Energy Systems Biology Knowledgebase (KBase, https://www.kbase.us/) is a good example of such computational infrastructure, which is designed to meet the grand challenge of predicting and designing biological functions. In addition to facilitating data access/sharing and building reusable bioinformatic pipelines, KBase uses a Narrative (an interactive digital notebook) to capture workflows for various scientific discoveries, which can be shared with other researchers to enhance scientific reproducibility and adaptability to answer other questions. The use of Jupyter Notebook-based narrative interface to encode workflows makes the computational framework much more accessible to the broader community. Another example is Pangeo (https://pangeo.io/), which is an open-source architecture that provides interconnected software packages and deployments of the software in cloud and high-performance computing environments for ocean, atmosphere, land and climate science. We will work with the ESS cyberinfrastructure working groups to collect the design requirements of the computational infrastructure from the broad community for maximum impact. Once built, it will provide transferrable scientific tools to understand watershed systems by iteratively learning from both process-based models, observational data, and data-driven approaches, paving our way towards a hybrid modelling approach that couples physical process models with the versatility of data-driven ML to improve the predictability of watershed models and Earth system models[8].

Suggested Partners/Experts

Manil Maskey, NASA/ESDS

Markus Reichstein, Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

Chris Henry, Argonne National Laboratory

References

1. Chen, X., R.M. Lee, D. Dwivedi, K. Son, Y. Fang, X. Zhang, E. Graham, J. Stegen, J. B. Fisher, D. Moulton, and T.D. Scheibe (2020). Integrating Field Observations and Process-based Modeling to Predict Watershed Water Quality under Environmental Perturbations, *Journal of Hydrology,* 125762, https://doi.org/10.1016/j.jhydrol.2020.125762.
2. Maskey, M., H. Alemohammad, K. J. Murphy, and R. Ramachandran (2020), Advancing AI for Earth science: A data systems perspective, *Eos, 101,* https://doi.org/10.1029/2020EO151245.
3. François-Lavet V., P. Henderson; R. Islam; M. G. Bellemare and J. Pineau (2018). An Introduction to Deep Reinforcement Learning. *Now Foundations and Trends*, doi: 10.1561/2200000071.
4. Zhou, J., G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun (2018). Graph Neural Networks: A Review Of Methods And Applications. arXiv:1812.08434.
5. Moshe, Z., A. Metzger, G. Elidan, F. Kratzert, S. Nevo, and R. El-Yaniv (2020). Hydronets: Leveraging river structure for hydrologic modeling. arXiv preprint arXiv:2007.00595.
6. Gkioxari, G., Malik, J. and Johnson, J., 2019. Mesh R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9785-9795. https://research.fb.com/publications/mesh-r-cnn/
7. Dramsch, J.S. (2020). 70 Years Of Machine Learning In Geoscience In Review. arXiv preprint arXiv:2006.13311v3.
8. Reichstein, M., G. Camps-Valls, B. Stevens. *et al.* (2019). Deep learning and process understanding for data-driven Earth system science. *Nature,* 566, 195–204 , https://doi.org/10.1038/s41586-019-0912-1.