Integrated parameter and process learning for hydrologic and biogeochemical modules in Earth System Models

2. Authors/Affiliations

Chaopeng Shen, Pennsylvania State University
Forrest M. Hoffman, Oak Ridge National Laboratory
Chonggang Xu, Los Alamos National Laboratory

3. Focus area

Primary focal area #2; secondary focal area #3: Learning about parameters and processes of land surface hydrologic and biogeochemical models in Earth System models by integrating machine learning, physics, and big data.

4. Science challenges

How do we maximally leverage big-data observations to improve hydrobiogeochemical process description and parameterization so that such modules more realistically capture hydrologic and vegetation responses and feedbacks under the future climate? For example, how can we leverage physics, limited observations of vegetation and streamflow to better estimate evapotranspiration, and, relatedly, net primary productivity, especially for drought areas? Vegetation plays a critical role in regional and global water cycles; however, existing vegetation models have failed to predict vegetation response to droughts (McDowell & Xu, 2017), arctic greening (Keenan & Riley, 2018), and critical transitions between forest and savanna (Hirota et al., 2011). These studies suggest that when we build process-based models (PBM) parameterized from regional and global plant traits, we tend to poorly describe plant adaptation and local-scale competition processes. The models and their associated parameters assigned for different regions in the world are not capturing essential heterogeneity in vegetation responses at finer spatial scales.

Many parameters of the land surface models control hydrology and vegetation dynamics at the same time. The heterogeneity in vegetation response is a function of (i) plant type, (ii) plant size, (iii) competition and succession, (iv) environmental controls, and (v) local variations due to the unique ecological community that are very difficult to describe (e.g., the size of gaps resulting from fire that facilitated the coexistence of pioneering species). In the demographic models, only factors (i) and (iv) were captured, and plant types were generally described only by leaf phenology and climate zones. With current demographic models, we generally consider more traits to define plant types (i) and calibrate these traits to consider factors (ii), (iii) and (iv); however, it is substantially challenging to scale to regional and global simulations due to trait variations across space (Ali et al., 2016). Moreover, it has been noted that hillslope processes, including ridge-to-valley flow and sunny vs. shady slopes are primary organizers of water, energy, and vegetation (Clark et al., 2015; Fan et al., 2019). Although gradual improvements in the hydrologic model component in earth system models may reduce this error (at a remarkably slow pace), the long-term, gradual impact of hydrology on plant traits are not well captured. Recent work showed that the hydrologic controls exerted by groundwater and lateral flow are primary regulators of rooting depth (Fan et al., 2017). Such hydrologic controls have seldom been reflected in vegetation model parameterizations.

5. Rationale

DOI: https://doi.org/10.2172/1769724
DOE BER, AI4ESP 2021 #AI4ESP1121 (https://www.ai4esp.org)

While purely data-driven deep learning techniques, partly led by authors of this whitepaper, have proven to be extremely powerful in hydrologic applications (Shen, 2018; Shen et al., 2018), especially in modeling soil moisture (Fang et al., 2017, 2019; Fang & Shen, 2020), streamflow (floods) (Feng et al., 2020), snow (Meyal et al., 2020), and water quality indicators like water temperature (Rahmani et al., 2020) and dissolved oxygen (Zhi et al., 2020), they are constrained by data availability and cannot make predictions in variables that are not directly observed at large scales, e.g., groundwater flow and evapotranspiration (there are global satellite-based estimates, but they are not direct and contain substantial modeled elements; there are also in-situ data at hundreds of sites, but they are far from covering the heterogeneity of the world). As mentioned earlier, how can we utilize multifaceted observations to inform parts of the water cycle that is not observed?

A primary paradigm to describe vegetation heterogeneity, compensate for model errors and produce more realistic outputs is to utilize model physics and calibrate the model parameters with observations. However, traditional parameter calibration has been a process that is fundamentally inefficient and ambiguous. The present generation of evolutionary, genetic, or similar algorithms (Deb et al., 2002; Maier et al., 2014; Price et al., 2005) requires thousands of model runs, if not more, to calibrate a dozen parameters. The evolutionary algorithms (EAs) optimize by trial and error across many discretize generations, involving many random walks. These methods are normally quite computationally expensive and are difficult to be applied for the global dynamic vegetation models.

The parameter calibration paradigm also faces the well-known problem of parameter non-uniqueness (referred to by some as equifinality (Beven, 2006; Tang & Zhuang, 2008)). Many parameters serve similar purposes in a model and are under-constrained. Parameter regionalization approaches may mitigate the issue but their performance may be sub-optimal. Addressing these issues may require us to fundamentally alter the way we think about the procedure for parameter (and process) determination. It may also require the addition of many constraints, which may not be convenient to do with EAs. Also we need to assess parameter uncertainty, often using Markov Chain Monte Carlo (MCMC)-type methods. MCMCs guess at the parameter distributions and gradually improve the guess with iterations. They are even more expensive than calibrations and are challenging to run at earth system scales.

The ineffectiveness of the present parameter calibration paradigm is entangled with process uncertainty, data noise and various other issues. Apart from being a perennial strain on computing resources and human labor, it is also one of the fundamental barriers to progress in earth system modeling.

6. Narrative

The recent progress in machine learning (ML) has opened up an avenue toward learning about processes and parameters for either existing or new process-based land surface models (PBM). The easiest point of entry may be model parameters but there may be many opportunities for formulation improvement as we proceed. Once we are able to improve parameterization and model structures using an integrated ML-PBM approach, we may be able to reduce computational demand by orders of magnitude, and obtain more sensible parameters/processes. Potential approaches include “parameter learning” (Tsai et al., 2020). The new generation of parameter learning approach does not just use neural networks for training a surrogate model -- it

organically couples physical descriptions in existing numerical models with neural networks and turns the parameter calibration into a machine learning problem to truly exploit the value of data. Neural-network-based methods like this are also super efficient -- parameter learning was able to reduce the computational demand by four orders of magnitude compared to the conventional paradigm and this savings is far beyond what the savings from a surrogate model would offer. By advancing techniques like this, we should then be able to:
(i) Make parameter estimation a much more trivial task, allowing more thorough examination to occur at different levels of complexity.
(ii) Develop a systematic parameter estimation tool for a variety of land surface hydrologic models, including the E3SM land model (ELM), ATS, PFLOTRAN, etc.
(ii) Use parameters as "process probes" to understand how to better improve the models and, based on data, identify sub-program units where current formulations are lacking.
(iii) More easily quantifying uncertainties as computational demand is reduced by orders of magnitude.
(iv) Compared to data-driven models, better predict extreme scenarios and trends.

These integrated ML-PBM approaches may allow us to answer many transformational questions to explore along these paths which may be difficult to answer otherwise, including:

(1) How do soil, topography, and groundwater influence plant hydraulic and root distribution parameters? What are the relative importance of topography, environmental variables, and uncaptured heterogeneity in determining plant traits that control ET and GPP during drought, e.g., target carbon storage and stomata closure points?
(2) How to best achieve parameter scaling for processes controlled by hydrology, which is known to exhibit scale-dependence?
(3) How can spatially-integrated observations like streamflow inform grid-scale hydrologic processes? How to best extrapolate information available from limited in-situ studies to global-scale studies?
(4) How strong do local variations due to landscape positioning and hydrologic control vegetation trait parameters?
(5) Does better parameter representation of vegetation also lead to better hydrologic simulations, in both low flow and high flow ranges?
(6) What are the relevant constraints for hydrologic and vegetation dynamics? More constraints means a reduction in uncertainty. We want to explore the possible constraints that could be added and the corresponding physics. We also need a neural-network-native way of quantifying different sources of uncertainty.
(7) How to shed light on process improvements? The parameters could serve as "process probes" or "handles" into the model dynamics and their discrepancies with the observations. What kind of diagnosis activities can we use to examine and revise some of the basic assumptions in ELM?

We hope to start the discussion by focusing on parameters and processes controlling evapotranspiration which couples the water cycle and vegetation feedbacks to the atmosphere; Then we discuss how to better constrain runoff based on basin-integrated streamflow observations at limited sites, given multifaceted data sources; We want to further extrapolate the value of data to unobserved variables and processes.

7. Suggested partners/experts

The main authors are Chaopeng Shen (PSU), Forrest M. Hoffman (ORNL), Chonggang Xu (LANL). Shen already has led foundational work in parameter learning, which could be shared with the community.

Collaborators include Daniel Kifer, Computer Science and Engineering (PSU), but we imagine there will be many collaborators.

8. References

Ali, A. A., Xu, C., Rogers, A., Fisher, R. A., Wullschleger, S. D., Massoud, E. C., Vrugt, J. A., Muss, J. D., McDowell, N. G., Fisher, J. B., Reich, P. B., & Wilson, C. J. (2016). A global scale mechanistic model of photosynthetic capacity (LUNA V1.0). *Geoscientific Model Development*, *9*(2), 587–606. https://doi.org/10/f8qhr3

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10/ccx2ks

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R. P., Kumar, M., Leung, L. R., Mackay, D. S., Maxwell, R. M., Shen, C., Swenson, S. C., & Zeng, X. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, *51*(8), 5929–5956. https://doi.org/10/f7wc44

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10/bnw2vv

Fan, Y., Clark, M., Lawrence, D. M., Swenson, S., Band, L. E., Brantley, S. L., Brooks, P. D., Dietrich, W. E., Flores, A., Grant, G., Kirchner, J. W., Mackay, D. S., McDonnell, J. J., Milly, P. C. D., Sullivan, P. L., Tague, C., Ajami, H., Chaney, N., Hartmann, A., … Yamazaki, D. (2019). Hillslope hydrology in global change research and earth system modeling. *Water Resources Research*, *55*(2), 1737–1772. https://doi.org/10/gfv8bz

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., & Otero-Casal, C. (2017). Hydrologic regulation of plant rooting depth. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(40), 10572–10577. https://doi.org/10/gb26pv

Fang, K., Pan, M., & Shen, C. (2019). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2221–2233. https://doi.org/10/gghp3v

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *21*(3), 399–413. https://doi.org/10/ggj669

Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophysical Research Letters*, *44*(21), 11,030-11,039. https://doi.org/10/gcr7mq

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019WR026793

Hirota, M., Holmgren, M., Van Nes, E. H., & Scheffer, M. (2011). Global resilience of tropical forest and savanna to critical transitions. *Science (New York, N.Y.)*, *334*(6053), 232–235. https://doi.org/10/fvr3cn

Keenan, T. F., & Riley, W. J. (2018). Greening of the land surface in the world's cold regions consistent with recent warming. *Nature Climate Change*, *8*(9), 825–828. https://doi.org/10/gd6hpx

Maier, H. R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L. S., Cunha, M. C., Dandy, G. C., Gibbs, M. S., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D. P., Vrugt, J. A., Zecchin, A. C., Minsker, B. S., Barbour, E. J., Kuczera, G., Pasha, F., … Reed, P. M. (2014). Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. *Environmental Modelling & Software*, *62*, 271–299. https://doi.org/10/f6tw49

McDowell, N. G., & Xu, C. (2017). Using traits to uncover tropical forest function. *New Phytologist*, *214*(3), 903–904. https://doi.org/10/ggj68c

Meyal, A. Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., & Wainwright, H. (2020). Automated cloud based long short-term memory neural network based SWE prediction. *Frontiers in Water*, *2*. https://doi.org/10.3389/frwa.2020.574917

Price, K., Storn, RM., & Lampinen, JA. (2005). *Differential evolution: A practical approach to global optimization*. Springer-Verlag.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2020). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/abd501

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10/gd8cqb

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X., & Tsai, W.-P. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Tang, J., & Zhuang, Q. (2008). Equifinality in parameterization of process-based biogeochemistry models: A significant uncertainty source to the estimation of regional carbon dynamics. *Journal of Geophysical Research*, *113*(G4), G04010. https://doi.org/10/btw8qf

Tsai, W.-P., Pan, M., Lawson, K., Liu, J., Feng, D., & Shen, C. (2020). From parameter calibration to parameter learning: Revolutionizing large-scale geoscientific modeling with big data. *ArXiv:2007.15751 [Preprint]*. http://arxiv.org/abs/2007.15751

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., & Li, L. (2020). *From hydrometeorology to water quality: Can a deep learning model learn the dynamics of dissolved oxygen at the continental scale?* [Preprint]. https://doi.org/10.1002/essoar.10504429.1