Modular hybrid modeling to increase efficiency, explore structural uncertainty, and allow multidimensional complexity scaling in land surface models.

C. Koven[1], R. Knox[1], R. Fisher[2], F. Hoffman[3], T. Keenan[1], D. Lawrence[4], M. Longo[1,5], B. Sanderson[6]
[1]Lawrence Berkeley National Lab, Berkeley, CA, USA; [2]Centre National de la Recherche Scientifique, Toulouse, France; [3]Oak Ridge National Lab, Oak Ridge TN, USA; [4]National Center for Atmospheric Research, Boulder, CO, USA; [5]Jet Propulsion Lab, Pasadena, CA, USA; [6]Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, Toulouse, France

Focal Area 2: Predictive modeling through the use of AI techniques and AI-derived model components

Science Challenge: Land surface models (LSMs) are indispensable tools for predicting hydrologic extremes, as well as a particularly uncertain component of Earth system models that has stubbornly resisted the convergence in projections over several successive generations of model intercomparisons. This uncertainty in LSMs is poorly quantified and poorly attributed to specific processes, which has hampered efforts to focus research in reducing uncertainty. This has resulted from sparse sampling of the possible uncertainty space—which is high-dimensional and has contributions from parametric, structural, initial, and boundary condition uncertainties—as an artifact of CMIP-type ensembles of opportunity and limitations inherent in observational benchmarks. A new approach is needed to understand and reduce this uncertainty, based around individual LSMs that can represent the breadth of assumptions represented in current CMIP-type efforts, while at the same time exploring that uncertainty in a systematic way, confronting multiple types of observations, and where justified, replacing process representations with ML-driven emulators. We propose an approach of modular hybrid modeling to address these challenges.

Rationale: Land surface modeling for projection of hydrologic extremes and carbon cycle dynamics is hampered by wide model disagreement. This arises from a number of sources, including: the high spatial heterogeneity of the land surface and the associated computational cost of resolving this heterogeneity, high parametric uncertainty, a high degree of structural variation in models that is itself reflective of a high degree of epistemic uncertainty in process representation, and a large number of interacting processes. To meet these challenges, a new generation of LSMs must be structured around modular representation of processes, with a focus on the ability to substitute a wide variety of processes with both multiple direct process representations and AI-based emulators—of either process models or direct observations. Such an approach, shown in Figure 1, will allow scientific progress through three distinct pathways: (1) the substitution of computationally complex model components, such as the coupled leaf photosynthetic and transpiration pathways, with ML-based emulators, will enable larger ensembles and greater resolution of spatial heterogeneity through increased computational efficiency and the ability to exploit novel HPC machine architectures; (2) the substitution of processes with a high degree of epistemic uncertainty with ML-driven emulators of either direct observations or multiple structurally different processes will enable greater model fidelity and ability to explore structural uncertainty; (3) the ability to separate large subsets of processes within LSMs and replace them with ML-based emulators will enable the separation of internal feedbacks and a wider degree of applications to different domains, spatial datasets.

DOI: https://doi.org/10.2172/1769750
DOE BER, AI4ESP 2021 #AI4ESP1071 (https://www.ai4esp.org)

Narrative:
Land surface models are indispensable tools for projecting the dynamics of water, carbon, disturbance, and other key aspects of the global system under the combined pressures of climate change, elevated CO_2, and changing human management of the Earth's lands. However, they also represent one of the least well constrained aspects of the Earth system. Carbon cycle feedback estimates from the land surface have high uncertainty that has not appreciably narrowed across several generations of Earth system models, including most recently the CMIP6 ensemble (Arora et al. 2020). Land surface models are currently used to probe diverse questions, such as the role of the land in governing hydrologic extremes, through processes such as 1) CO_2 fertilization and its impacts on flooding risk (Kooperman et al. 2018), 2) anthropogenic disturbance and its control of ecosystem functioning (Longo et al. 2020) to 3) coupled land-atmosphere hydrologic functioning (Knox et al. 2015). Representation of the large set of interacting processes required to span these questions results in a high degree of model complexity and uncertainty. A much more systematic and detailed characterization of model uncertainty is required to improve predictions and projections in the Earth's coupled carbon and hydrological systems.

To achieve this goal of improved prediction with LSMs, hybrid modeling approaches that allow replacement of subsets of process-based models will enable progress to be made along at least three separate problems. The first of these is that a small set of individual processes, which account for a disproportionate share of model computational cost, could be replaced with machine-learning-based emulators of a given process-based parameterization. Examples include, in particular, the coupled photosynthesis, stomatal conductance, and canopy turbulence equations, which must be solved iteratively every timestep and account for a high fraction of model cost. Because these processes are governed by plant traits that vary widely between co-occurring plants and across hillslope, disturbance, and other fine-scale gradients, reducing the cost of these equations will permit much greater model fidelity with respect to the heterogeneity of the land surface, and at the same time allow much greater representation of functionally diverse plants, by allowing greater numbers of co-occurring plant functional types, thereby allowing better resolution of ecological dynamics through competition of functionally diverse communities. Achieving this will require novel methods that allow predicting when emulator errors require reverting to conventional parameterizations.

A second problem for which modular hybrid modeling could enable rapid progress is the representation of processes with particularly large degree of model structural uncertainty, but where large datasets exist that are comprehensive enough to avoid extrapolation errors and could be used to train empirical/ML models. Examples of this may include phenology or other processes with large amounts of remotely-sensed data, or plant allometry, given the rapid growth of lidar-derived datasets. Because several such processes—characterized by a low degree of fidelity and low mechanistic justification—exist in current ESMs, replacing them with more sophisticated empirically-derived models may allow for greater overall model fidelity. Further, even where observational training datasets do not exist, there may be opportunities to find continuous approximations of structurally-diverse process parameterizations in LSMs and thereby shift structural to parametric uncertainty, as has been demonstrated in GCMs (Lambert et al. 2020).

A third problem that a modular hybrid modeling system may solve is in the scaling of model complexity. The high degree of process coupling results in high complexity of LSMs, which presents challenges both to model understanding and in isolating specific process uncertainties. Thus there is always a need for simplified configurations of LSMs. However,

there are many different examples of a simplified LSM, because appropriate simplifications are themselves dependent on the context of the specific problem being addressed (Fig 1b). Our proposed simplified configurations essentially depend on isolating a set of processes and excluding the set of external processes. If models were constructed with clearly defined boundaries and interfaces, then it would be much more straightforward to force any given subset of the land surface processes with an emulated representation of the boundary conditions from the full model. Doing so would widen the set of use cases and allow isolation of process uncertainty, while also creating the potential for tractable machine learning component emulation.

We propose that novel model architectures are required to facilitate the use of modular hybrid modeling approaches. These will require a more coordinated approach of isolating specific processes, with clearly defined boundaries, rather than the more haphazard way in which models have historically been built. Starting with the most computationally expensive parts of LSMs, we could build out to isolate distinct submodels in ways that allow the replacement of individual parts or combinations of parts with machine-learning-based emulators. Aspects of the open-source FATES model, with more modular photosynthesis representation as well as separation of the entire demographic aspect from the rest of the ELM, may provide a template for further development along these lines.

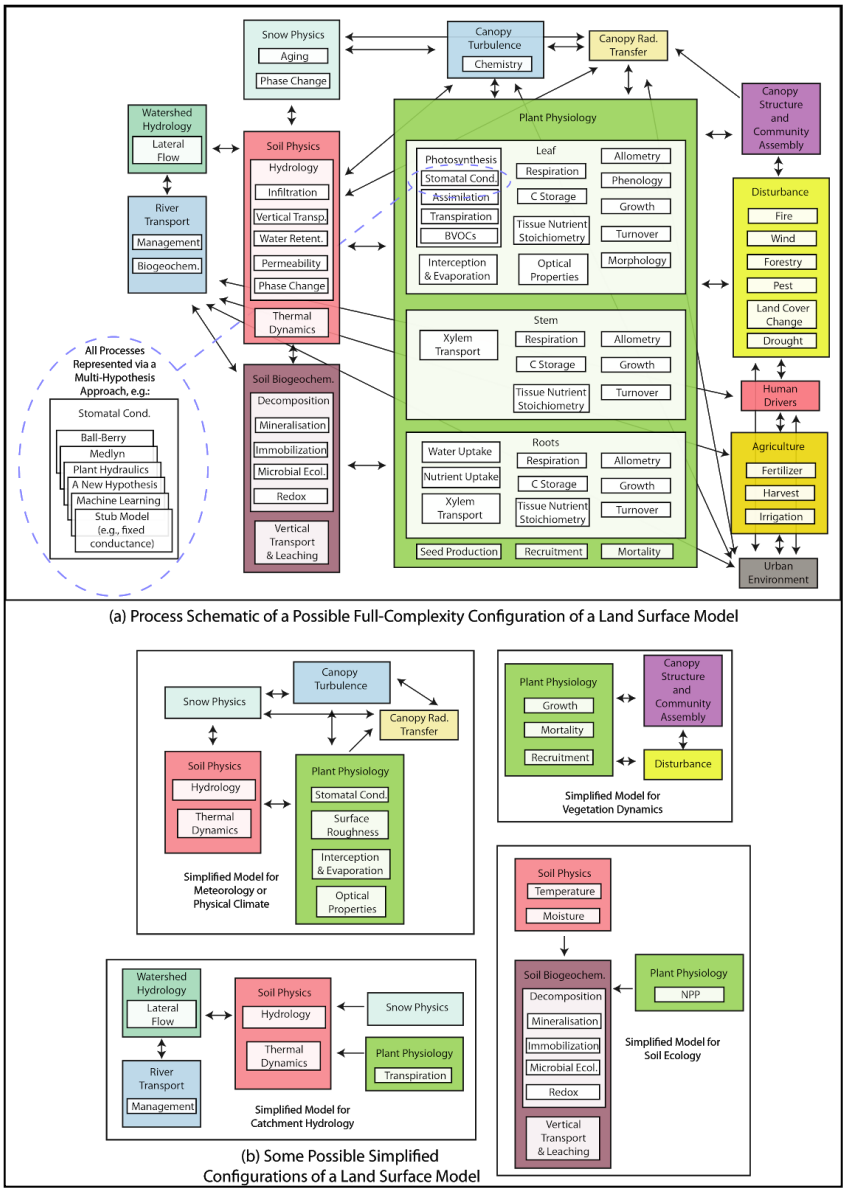

Figure 1 Schematic of how modular hybrid modeling could work in LSMs such as FATES. (a) Specific processes such as stomatal conductance could be replaced by ML-based emulators to enable computational speedup; (b) Larger subsets of models could be replaced with ML-based emulators to enable much more flexible configuration of LSMs to address a wider variety of specific problems. From (Fisher and Koven 2020).

References

Arora, Vivek K., Anna Katavouta, Richard G. Williams, Chris D. Jones, Victor Brovkin, Pierre Friedlingstein, Jörg Schwinger, et al. 2020. "Carbon–concentration and Carbon–climate Feedbacks in CMIP6 Models and Their Comparison to CMIP5 Models." https://doi.org/10.5194/bg-17-4173-2020.

Fisher, Rosie A., and Charles D. Koven. 2020. "Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems." *Journal of Advances in Modeling Earth Systems* 12 (4). https://doi.org/10.1029/2018MS001453.

Knox, Ryan G., Marcos Longo, Abigail L. S. Swann, Ke Zhang, Naomi M. Levine, Paul R. Moorcroft, and Rafael L. Bras. 2015. "Hydrometeorological Effects of Historical Land-Conversion in an Ecosystem-Atmosphere Model of Northern South America." *Hydrology & Earth System Sciences* 19 (1). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.904.6427&rep=rep1&type=pdf

Kooperman, Gabriel J., Megan D. Fowler, Forrest M. Hoffman, Charles D. Koven, Keith Lindsay, Michael S. Pritchard, Abigail L. S. Swann, and James T. Randerson. 2018. "Plant Physiological Responses to Rising CO2 Modify Simulated Daily Runoff Intensity With Implications for Global-Scale Flood Risk Assessment." *Geophysical Research Letters* 45 (22): 12–457.

Lambert, F. H., P. G. Challenor, N. T. Lewis, D. J. McNeall, N. Owen, I. A. Boutle, H. M. Christensen, et al. 2020. "Continuous Structural Parameterization: A Proposed Method for Representing Different Model Parameterizations within One Structure Demonstrated for Atmospheric Convection." *Journal of Advances in Modeling Earth Systems* 12 (8): e2020MS002085.

Longo, M., S. S. Saatchi, M. Keller, K. W. Bowman, A. Ferraz, P. R. Moorcroft, D. Morton, D. Bonal, P. Brando, B. Burban, G. Derroire, M. N. dos-Santos, V. Meyer, S. R. Saleska, S. Trumbore, and G. Vincent. 2020. "Impacts of degradation on water, energy, and carbon cycling of the Amazon tropical forests." *J. Geophys. Res.-Biogeosci.*, 125 (8), e2020JG005677, https://doi.org/10.1029/2020JG005677.