Using machine learning to improve land use/cover characterization and projection for scenario-based global modeling.

Alan Di Vittorio (LBNL) – lead, Kate Calvin (PNNL)

Focal Areas

This proposal focuses on both (1) data acquisition/assimilation and (2) predictive modeling. In the case of land data, historical reconstruction requires both elements due to sparse temporal data, and future projection is dependent upon the initial state and historical trends.

Science Challenge

The characterization of the land surface in models is critical for robust estimation of the water cycle and associated extremes. Land use determines water demand, and land cover affects water availability. The interaction between water demand and availability is the key determinant of whether systems are resilient to variability in the overall water cycle. Furthermore, activities such as irrigation (DeAngelis et al., 2010) and deforestation or afforestation may influence regional (Debortoli et al., 2016) to global (Swann et al., 2012) precipitation patterns, depending on the extent of the activity. Thus this proposed work aims to address the following question: How does integrated land use/cover data and improved land use/cover projection, as informed by machine learning approaches, better resolve human-earth system resilience to water cycle variability and extremes?

Rationale

There is considerable uncertainty in both historical and future projections of land use/cover that has not been characterized but directly affects climate, carbon, and hydrological projections in global economic, Multisector Dynamics, and Earth system models. The substantial amount of data and processing required to generate the initial and bounding land data, combined with rigid model structures, leads to a single, and often unique, Earth being simulated by each global model. Due to large time and effort requirements, each modeling group selects a limited number of data sources and creates a single representation of the land surface. Furthermore, there is a false dichotomy between land cover and land use that is currently implemented by separate processing and incomplete or inconsistent integration of the two into a whole land surface for modeling. These challenges are even more apparent when economic models produce land scenarios that are incompatible with the Earth system models they are intended to drive. Additionally, global land scenario generation is limited to simple economic constraints, while agent-based alternatives are largely unconstrained and limited to regional applications due to computational costs.

Machine learning has the potential to improve land characterization through better integration of multiple data sources with explicit quantification of uncertainty. Advanced pattern recognition can provide a set of more reasonable historical reconstructions than traditional rule-based back-projection techniques, largely by incorporating geographic

DOI: https://doi.org/10.2172/1769796

heterogeneity and multiple data sources instead of applying a single, global rule to one data source.

Machine learning also has the potential to improve land use/cover projection by incorporating multiple factors into land use decisions based on observed patterns. This extends economic projection in a more constrained way, and with the potential for better computational performance, than complex, rule-based many-agent models.

The application of machine learning to land data will facilitate two major advances in global modeling in the next 10 years: 1) a paradigm shift from separate land use and land cover to an integrated and evolving land surface, and 2) a new form of land scenario projection that bridges the gap between simple economics and complex, unconstrained agents.

Narrative

We propose to incorporate biogeophysical and socio-economic data into machine learning approaches to both generate historical reconstructions of and project future global land use/cover. There are a few different machine learning approaches that may be viable for this (e.g., Reichstein et al., 2019; Huntingford et al., 2019), but currently most rely on geophysical data and are regional in extent (e.g., Abdullah et al., 2019; Jin et al., 2018), which inhibits their application to future projection. Some have incorporated multiple data sources to capture input uncertainty (e.g., Alemohammad et al., 2017), which is a valuable way to quantify uncertainty associated with the various land cover data available, and can be used to generate multiple data sets to represent this and other data uncertainties. We intend to take advantage of machine learning algorithms, advanced DOE computing resources, and a multitude of global land data to advance land science in support of Earth system prediction by providing integrated land data sets for use by different types of models such as GCAM and E3SM.

Developing multiple approaches may be effective in addressing different needs such as providing initial/bounding conditions, projecting land use/cover within the context of another model, and model assessment. For example, artificial neural networks (e.g., Scott et al., 2017) and random forest algorithms (e.g., Jin et al., 2018) have been successful at classifying various types of land, and can be applied to generate initial conditions. Further development of neural networks, such as the convolutional LSTM network (Shi et al., 2015), allows for the temporal projection of spatial data and could be applied to both historical reconstruction (initial and bounding conditions, model evaluation) and future projection of land use/cover (for assessing other models). In the context of scenario-based global modeling, where economic models project future land use/cover for ESMs, the historical land use/cover reconstruction can be performed with one method and the results applied to historical ESM projections and economic model initializations for consistency. To improve the reconstruction and facilitate the transition from historical to future projection, socio-economic data such as land tenure, land value, commodity production, and operational costs of land use could be incorporated into the processing.

Incorporating machine learning into the land projection within an economic model has the potential to expand the model's scope beyond simple economics and can be done in a variety of ways. One option is to train a convolutional LSTM network on historical data corresponding with the information available within an economic model (e.g., GCAM) and

insert it directly into the economic model. Another possibility is to use machine learning and observational data to constrain and develop a reduced order representation of an agent-based or statistical land projection model and use this representation within the economic model. Depending on the algorithm, some driving variables may be represented as distributions and their thresholds determined during training, such as for extremely randomized decision trees (Geurts et al., 2006).

The proposed work directly addresses three of the Data-Model Integration research goals in support of three scientific grand challenges (Integrated Water Cycle, Biogeochemistry, and Drivers and Responses in the Earth System), with direct implications for Integrated Water Cycle research. 1) The innovative tools and data products resulting from this work can be used to both drive and evaluate Earth system models (ESMs) and will also help understand and characterize uncertainty in land data drivers and its contribution to uncertainty in ESM simulations. 2) This work will facilitate development of ESMs that effectively utilize observational land constraints and allow for uncertainty in ESM outputs to guide further development of input land data. 3) While geared toward data integration and projection that will enable land data consistency across DOE's global models (e.g., GCAM, E3SM), the global data could be regionally subset and the proposed approaches could be applied to data at various resolutions and extents to facilitate interoperability across temporal and spatial scales.

While these Data-Model Integration goals apply to three scientific grand challenges, the proposed work focuses on how they address three research goals of the Integrated Water Cycle scientific grand challenge. 1) How does changing land use/cover affect the hydrological functioning of watersheds and river basins? 2) How do energy, water, and land systems co-evolve in response to perturbations? 3) To what extent do local variations in land use/cover drive larger-scale hydrological processes?

The proposed tools and data required to address these research goals will be reusable and the findings reproducible. All code and data will be assigned an open source license. Data products will also be released as citable archives, and we will promote evaluation and use of these products and tools through funded modeling projects (e.g., GCAM, E3SM) and professional international groups such as the Future Earth Analysis, Integration, and Modeling of the Earth System working group on Human-Earth System Modeling. Our plans to make the proposed tools and data publicly available align with FAIR principles. Code, metadata, and data will all be assigned unique digital object identifiers and be archived in a searchable database such as zenodo.org or the new Multi-Sector Dynamics data server. They will be easily retrievable by manual or automated protocols (e.g., http, ftp) and thoroughly described using accepted vocabularies.

The overall goal of the proposed work is to advance Earth system science by providing improved data and projections of land use/cover using novel tools and working to ensure that these products are utilized by the broader Earth system science community.

References

Abdullah, A.Y., A. Masrur, M. Sarfaraz, G. Adnan, M.A.A. Baky, Q.K. Hassan, and A. Dewan. Spatio-temporal patterns of land use/land cover change in the heterogeneous coastal regions of Bangladesh between 1990 and 2017 (2019). Remote Sensing, 11:790.

Alemohammad, S.H., B. Fang, A.G., Konings, F. Aires, J.K. Green, J. Kolassa, D. Miralles, C. Prigent, and P. Gentine (2017). Water, energy, and carbon with aritificial neural networks (WECANN): a statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence. Biogeosciences, 14:4101-4124.

DeAngelis, A., F. Dominguez, Y. Fan, A. Robock, M.D. Kustu, and D. Robinson (2010). Evidence of enhanced precipitation due to irrigation over the Great Plains of the United States. Journal of Geophyscial Research Atmospheres, 115:D15115.

Debortoli, N.S., B. Dubreuil, M. Hirota, S.R. Filho, D.P. Lindoso, and J. Nabucet (2017). Detecting deforestation impacts in Southern Amazonia rainfall using rain gauges. International Journal of Climatology, 37:2889-2900.

Geurts, P., D. Ernst, L. Wehenkel (2006). Extremely randomized trees. Machine learning, doi: 10.1007/s10994-006-6226-1.

Huntingford, C., E.S. Jeffers, M.B. Bonsall, H.M. Christensen, T. Lees, and H. Yang (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14:124007.

Jin, Y., X. Liu, Y. Chen, and X. Liang (2018). Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: a case study of central Shandong. International Journal of Remote Sensing, doi: 10.1080/01431161.2018.1490976.

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

Scott, G.J., M.R. England, W.A. Starms, R.A. Marcum, and C.H. Davis (2017). Training deep convolutions neural networks for land-cover classification of high-resolution imagery. IEEE Geoscience and remote sensing letters, 14(4):549-553.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo (2015). Deep learning for precipitation nowcasting: a benchmark and a new model. *Adv. Neural. Inf. Process. Syst*. **30**, 5617–5627.

Swann, A.L.S., I.Y. Fung, and J.C.H. Chiang (2011). Mid-latitude afforestation shifts general circulation and tropical precipitation. Proceedings of the National Academy of Sciences, 109(3):712-176.