Using Statistical Learning Methods to Accelerate Model Parameter Sensitivity Experiments

Shamik Bhattacharya[1], Forrest M. Hoffman[2], Bharat Sharma[2,3], Gaurab KC[2], Nathan Collier[2] , Min Xu[2], Michael Kelleher[2]

1.North Carolina State University, Raleigh, North Carolina 27606; 2. Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830; 3. Northeastern University, Boston, Massachusetts 02115

Introduction

- The Sixth Assessment Report of the Intergovernmental Panel on Climate Change concluded that climate warming is unequivocal and human influence on the climate system is evident.
- Sea-ice loss, sea-level rise, and atmospheric and oceanic warming have been driven by increases in greenhouse gasses, and the consequences of warming will get worse in the twenty-first century.
- To understand the predicted effects of climate change, better ecosystem representation in land models linked to Earth system models is required.
- A large number of land model simulations is needed to test the sensitivity of model predictions to parameters; however, running these simulations over the full grid is computationally infeasible.
- We applied cluster analysis to identify global ecoregions—regions that are relatively homogenous in terms of climate and carbon characteristics.
- We aggregated grid cells from a historical land-model simulation to define ecoregions at different levels of division and a representative grid cell for each region to use in parameter perturbation experiments.
- We also assessed the bias brought about by the clustering simplification at various levels in order to guide the selection of ecoregion maps for upcoming simulation tests.

Materials and Methods

- In our first analysis, we used the annual mean of 19 variables from a land model simulation to characterize the land drivers and hydrological and biogeochemical responses averaged over each of six 25-year windows from 1858 to 2008.
- In a second analysis we used means and standard deviations of the same 19 variables to characterize both the average and the interannual variability of those land drivers and responses in each 25-year window.
- We used a highly scalable, parallel *k*-means cluster analysis algorithm to produce ecoregions (roughly homogeneous land areas) from the mean and mean & sd data sets at many levels of division (*k*), where *k* is the number of groupings or clusters or ecoregions desired.
- We compared both the annual (mean) and annual & sd (mean and the standard deviation) representations of the original model output for each of the 19 variables using the The International Land Model Benchmarking (ILAMB) package (Collier et al., 2018) to determine a sufficient representation for the actual model.
- We used Python in a Jupyter Notebook to create maps of ecoregions and their most-representative grid cells.

Results

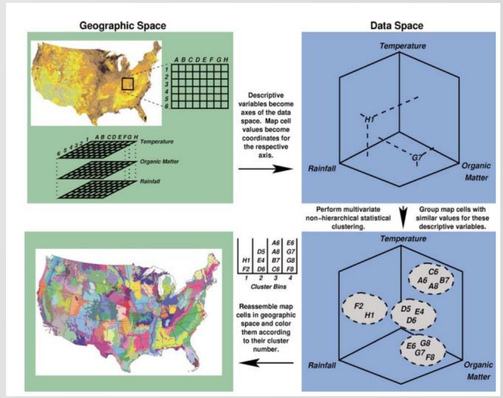

Figure 1: A transition from geographic space (left) to data space (right), an iterative k-means cluster analysis, and a transformation back to geographic space are all components of Multivariate Geographic Clustering (MGC).

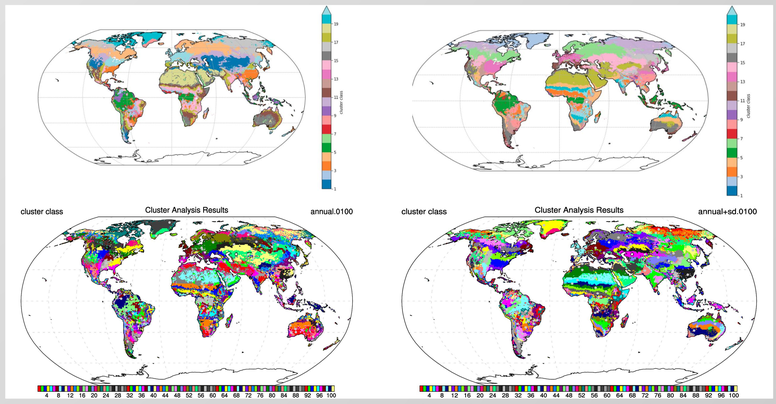

Figure 2: In the top row and bottom row of the image, respectively, are cluster class maps for k = 20 and k = 100 for the annual mean and standard deviation, and the realized centroid locations are shown by grey filled circles in the left and right columns, respectively. The maps get more detailed and show more of the structure of the original model output as the levels of division are increased by raising the value of k.

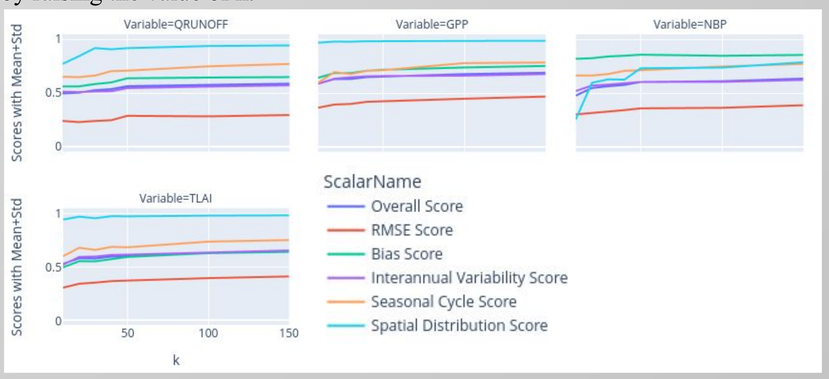

Figure 3: We evaluated the maps created solely utilizing the cluster realized centroid values using ILAMB. As the level of division (k) is increased, the representation scores for the several land variables (QRUNOFF, GPP, NBP, and TLAI shown above) rise. Beyond k=100, the scores do, however, improve more slowly.

Results (continued)

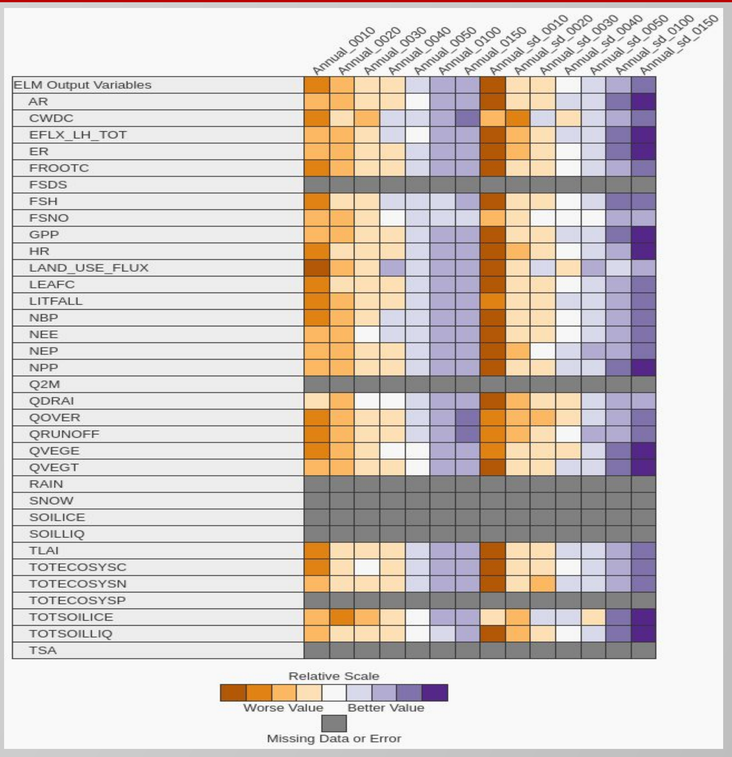

Figure 4: ILAMB shows the improvements in relative scores as the value of *k* is increased for both the annual means and the annual means & standard deviations.

Conclusions

- *k*-means cluster analysis provides a valuable method for sampling a large, multi-dimensional space of Earth system model results and determining locations that are representative of larger spatial areas.
- As the level of division (*k*) is increased, the number of ecoregions increases and the nominal size of the ecoregions decreases; thus, the regions become more specific or refined.
- ILAMB provides a useful way to quantitatively evaluate the biases from the model representation at each level of division.
- We are evaluating the accuracy of each clustered realization's representation of the original model output for the relevant key variables using ILAMB.
- A modeler may use this information to select a sufficient level of division and use the *k* realized centroid grid cells as representatives of ecoregions for modeling studies.

Literature Cited

- Collier, Nathan, Forrest M. Hoffman, David M. Lawrence, Gretchen Keppel-Aleks, Charles D. Koven, William J. Riley, Mingquan Mu, and James T. Randerson. November 1, 2018. “The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation.” *J. Adv. Model. Earth Syst.*, 10(11):2731–2754. doi:10.1029/2018MS001354.

Acknowledgments

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships program.