Ecological Cyberinfrastructure and HPC Towards More Accurately Predicting Future Levels of Greenhouse Gases

Rob Kooper[1], Kenton McHenry[1], Michael C. Dietze[2], David S. LeBauer[3], Shawn Serbin[4], Ankur Desai[4]
[1]National Center for Supercomputing Applications, University of Illinois
[2]Department of Earth and Environment, Boston University
[3]Energy Biosciences Institute, University of Illinois
[4]Atmospheric & Oceanic Sciences, University of Wisconsin-Madison
kooper@illinois.edu

Ecology research and other observational research fields are using computational models to predict future characteristics of complex systems. Here we consider ecology specifically and the prediction of plant and environmental traits over specified geographic locations and periods of time. Scientists have created a number of models such as the Ecosystem Demography (ED) model [4] and SIPNET (http://thesipnetmodel.blogspot.com/) which predict environmental attributes from plant growth to atmospheric carbon fluxes resulting from plant respiration. While all models are validated using real world observation they, like all scientific models, vary in terms of required input data, computational requirements, and overall accuracy. In addition to internal model assumptions, predictions are also affected by the quality of the input data provided to the model [2] and thus a good model will take these uncertainties into account and propagate them throughout each step in the model's workflow [1].

Towards providing scientists with better tools to study environmentally relevant questions, while also addressing the data and computational requirements of ecological modeling, we work to develop the Predictive Ecosystem Carbon Analyzer (PEcAn) [3]. Serving as cyberinfrastructure for ecology PEcAn incorporates a number of models while also providing interface adapters to relevant external data sources that can serve to generate needed model inputs. Data sources include (for a complete list see http://pecanproject.org/datasets) the FIA, NACP, the Biofuel Ecophysiological Trait and Yield database, NARR, and FAO. Work is underway to support more data formats such as MODIS, lidar, Palsar, Aviris, landsat. From these sources model input values are obtained for site information (plant types, characteristics, and density), weather information (temperature variations, rain fall, available light, wind speeds), soil information (sand, clay, etc.), and land use. A model's starting conditions are then created by sampling probabilistic distributions of these attributes. To analyze the uncertainty we create multiple runs using different initial values based on these probabilistic distributions. After all the runs are finished we create a probabilistic distribution for the outputs which gives us the uncertainty of the results. On top of this an easy to use web interface provides access to models, data, and execution on a variety of a computational resources. Once the model has finished execution we use datasets such as Ameriflux to verify the results of our complete system.

For our own studies we have been using ED primarily as it is the most accurate model currently available while not requiring significantly greater execution times compared to other models (minutes vs seconds). For a single site run over a single year the system requires 10 minutes of execution time on a typical desktop machine. For the uncertainty analysis we are looking at a 1000 runs for a single site. This process is repeated over multiple sites and will in the near future also include multi-year (up to century) runs, as well as regions (up to the whole of the continental US), each of which will increase the base computation time. Most HPC systems are not set up to handle the submission of thousands of relatively small jobs which is required for this type of analysis. We use a small MPI script to launch these thousands of sub-jobs and are currently evaluating performance on a number of XSEDE resources.

With the ease at which ecologists would have access to these complex computational models, and resources to run them on, we hope to allow for predictions across much longer time frames (i.e. from centuries to millennia). In addition this infrastructure will enable scientists to play what if scenarios, running models multiple times under varying initial conditions to see how things change.

ACKNOWLEDGMENTS

This research at UIUC, NCSA, and Boston University has been funded through the National Science Foundation Co- operative Agreement NSF DBI 10-62547.

REFERENCES

[1] Clark, J. S. 2007. *Models for ecological data: an introduction*. Princeton University Press Princeton.

[2] Dietze, M.C., LeBauer, D., and Kooper, R. 2012. *On improving the communication between models and data*. Plant, Cell & Environment.

[3] LeBauer, D., Wang, D., Richter, K., Davidson, C., and Dietze, M. C. 2012. *Facilitating feedbacks between field measurements and ecosystem models*. Ecological Monographs.

[4] Medvigy, D., Wofsy, S. C., Munger, J.W., Hollinger, D.Y., and Moorcroft, P.R. 2009. *Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem demography model version 2*. Journal of Geophysical Research: Biogeosciences.