2014 CDI Statement of Interest

SECTION 1. PROJECT ADMINISTRATIVE INFORMATION
CDI SSF Category 3: Data and Information Assets

Project Title: Species Distribution Modeling Using Cloud Computing Resources

USGS Cost Center: USGS Core Science Analysis and Synthesis

USGS Principal Investigator: Sky Bristol, Core Science Analysis and Synthesis, Denver, CO 80225. Tel: 303-241-4122 Email: sbristol@usgs.gov

Co-Investigators: Ole J. Mengshoel, Carnegie Mellon University, Silicon Valley Campus, Moffett Field, CA 94035. Tel: 650-335-2887, Email: ole.mengshoel@sv.cmu.edu

Collaborators: Derek Masaki, Core Science Analysis and Synthesis, 12201 Sunrise Valley Drive MS 302, Reston, VA 20192, Tel: 808-358-2497, Email: dmasaki@usgs.gov

Short Description: The current process of developing species distribution models is time consuming and requires that modelers be familiar with GIS, statistical methods, modeling packages, and other software. (For examples of species distribution modeling workflows, please see one using ArcGIS 10 [1] and another using R [2].) The output products from such workflows are typically static and require significant effort for each scenario. There is great potential in employing a simplified user interface, parallel processing methods, and moving the data to a high performance computing capability. This could dramatically simplify model processing for end users. A fast, interactive environment would provide a system encouraging non-modelers to run a range of scenarios to test a wide variety of potential inputs and responses. A realistic use case would involve: 1) moving species and landscape data (eg. amphibian species records and PRISM data) to a cloud computing platform; 2) developing a simple Bayesian model workflow with limited inputs and selections (eg. species>>geographic extent>>temperature parameter>>precipitation parameter); and 3) methods for generating a jpg or png image of the distribution map.

PROJECT SUMMARY

In order to improve the development of species distribution models, this research focuses on increasing the computational speed of Bayesian network and Markov random network inference. Using Gibbs sampling as the foundation, the research will enable faster processing of significantly larger data sets than what is currently feasible. We are planning to combine several proven methods, including chromatic Gibbs sampling and blocking Gibbs sampling, to develop faster algorithms. Our current approach is based on chromatic block Gibbs sampling, which first partitions a graph into a set of non-overlapping tightly connected node blocks, then samples blocks in parallel by taking advantage of their conditional independence properties.

The research consists of these major steps: 1) Developing our novel parallel Gibbs sampling algorithm (based on the ideas sketched above); 2) Implementing a prototype of the proposed algorithm in R; 3) Migrating the prototype into high performance platforms and languages (at the moment, the plan is to use a combination of Scala & GPU, based on GPU research performed by Prof. Mengshoel's group at CMU Silicon Valley); and 4) Performing tests and

experiments. The development and implementation of the R prototype will consist of the following minor steps: 1) Researching the best way to partition nodes into tightly connected components; 2) Implementing the junction tree algorithm in R (or find and adapt an existing implementation) to calibrate individual blocks, jointly sample all nodes in each block, as a substep in the chromatic block Gibbs sampling algorithm; 3) Implementing the proposed algorithm in an integrated fashion, and experiment with species distribution data sets.

The modeling effort will draw on species occurrence data provided through the USGS CSAS Biodiversity Information Serving Our Nation. A subset of 5 vascular plant species and 5 terrestrial vertebrate species will be used in the development of the modeling application.

The following deliverables will be provided: 1) Code base for the modeling workflow posted to USGS GitHub; 2) R-based prototype capable of being ported to a production server; 3) Distribution model outputs for 10 species.

[1] Nick Young, Lane Carter, and Paul Evangelista. "A MaxEnt Model v3.3.3e Tutorial (ArcGIS v10)," September 1, 2011. Natural Resource Ecology Laboratory at Colorado State University and the National Institute of Invasive Species Science.
[2] Robert J. Hijmans and Jane Elith. "Species distribution modeling with R." January 4, 2013.

BUDGET

Budget Category	Funding Requested	Matching Funds
1. SALARIES (including Benefits)		
Personnel Total	$10,000	$25,000
Contract Personnel Total	$15,000	$50,000
Total Salaries:	$25,000	$75,000
2. TRAVEL EXPENSES		
Travel Total:	$6,000	$6,000
Other travel expenses:	$0	$0
Total Travel Expenses:	$6,000	$6,000
3. OTHER DIRECT COSTS (itemize)		
Equipment (software, hardware)	$0	$25,000
Publication Costs	$0	$0
Other expenses	$0	$10,000
Total Other Direct Costs:	$5,000	$31,000
Total Direct Costs:	$36,000	$112,000
Indirect Costs (%)	15%	15%
Grand Total	**$48,300**	**$185,150**