# MIRAGE: A framework for data-driven collaborative high-resolution simulation

Byung H. Park[1], Melissa R.Allen[1], Devin White[1], Eric Weber[1], John T. Murphy[2],
Michael J. North[2], and Pam Sydelko[2]

[1]Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831
Telephone: (+1) 865.576.7658
Email: {parkbh, allenmr, whiteda1, weberem}@ornl.gov

[2]Argonne National Laboratory, 9700 S. Cass Avenue, Argonne IL 60439
Telephone: (+1) 630.252.2000
Email: {jtmurphy, north, psydelko}@anl.gov

## Abstract

Information about how human populations will shift in response to various stimuli is limited because no single model is capable of addressing these stimuli simultaneously, and integration of the best existing models has been challenging because of the vast disparity among constituent model purpose, architecture, scale and execution. To demonstrate a potential model coupling for approaching this problem, three major model components are integrated into a fully coupled system that executes a worldwide infection-infected routine where a human population requires a food source for sustenance and an infected population can spread the infection when they are in contact with the remaining healthy population. To enable high-resolution data-driven model federation and an ability to capture dynamics and behaviors of billions of humans, a high performance computing agent-based framework has been created and demonstrated.

## 1. Introduction

Various drivers—economic, environmental, technological, and social—will cause human populations to shift, changing the topology of urban infrastructure. These changes will create emerging vulnerabilities that, in anticipation of their consequences, will require integrated approaches and high-resolution data for local decision makers. Such information is currently limited, however, because no one model can address these drivers simultaneously, and integration of the best existing models has been challenging because of the vast disparity among constituent model purposes, architectures, scales and execution environments. The Foresight Initiative, a collaborative research effort led by the National Geospatial-Intelligence Agency and supported by Arizona State University and several national laboratories, is attempting to find solutions to these problems in order to bring the rich and complex world of computational modeling to bear on matters of national and international significance. One particular challenge of constructing such high-resolution data-driven model federations is to simulate billions of actors and their dynamics and behaviors in response to various scenarios. Thus it is imperative to bring high performance computing (HPC) capabilities into these studies. To demonstrate a possible method for addressing this challenge, we combined three major model components into a two-way-coupled system that executes a worldwide susceptible and infected (SI) epidemic simulation with a resource competition component built in—one that has the long-term potential to scale to HPC systems.

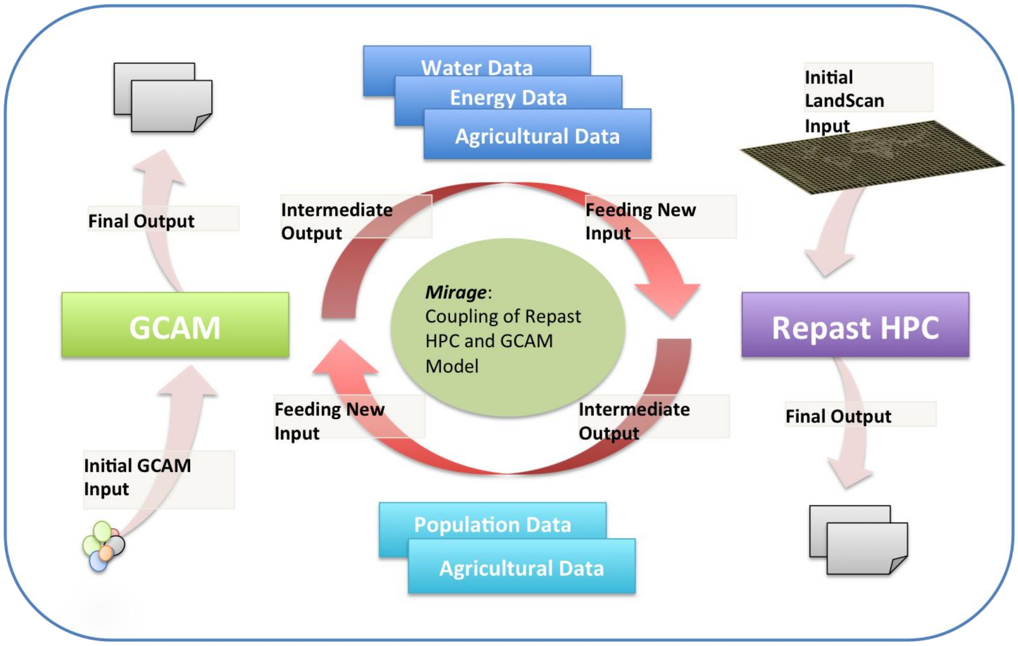


Figure 1. MIRAGE Framework.

A variety of methods for implementing the individual components of this set (Collier and North, 2012; McKee et al., 2015) have been implemented successfully, and some coupling for other purposes has been accomplished (e.g., Bond-Lamberty et al., 2014). Unfortunately, no individual model or modeling framework has fully addressed simultaneous influences among these particular components. In the sections that follow we introduce the *Modeling Interactive Repast and GCAM Ensemble* (MIRAGE) which is a framework for collaborative high-resolution simulations. We then detail the datasets, models, and procedures used for model intercommunication; the experiment that was performed using MIRAGE; the results that have been achieved to date; and our conclusions and recommended next steps.

## 2. Methods and Data

As shown in fig. 1, three components comprise MIRAGE: LandScan, GCAM, and a Repast HPC model. GCAM and the Repast model work in a closed loop, repeatedly feeding their outputs as inputs to the other model. Such a cycle of runs iterates until the results mature, when the final outputs are generated. The LandScan 2013 global population data set is the initial input to the Repast and GCAM models.

LandScan, at approximately 1 km resolution, was created using a multi-variable dasymetric modeling approach along with spatial data and imagery analysis technologies to disaggregate census counts within administrative boundaries at the town, county, and state levels (http://web.ornl.gov/sci/landscan/ landscan_documentation.shtml).

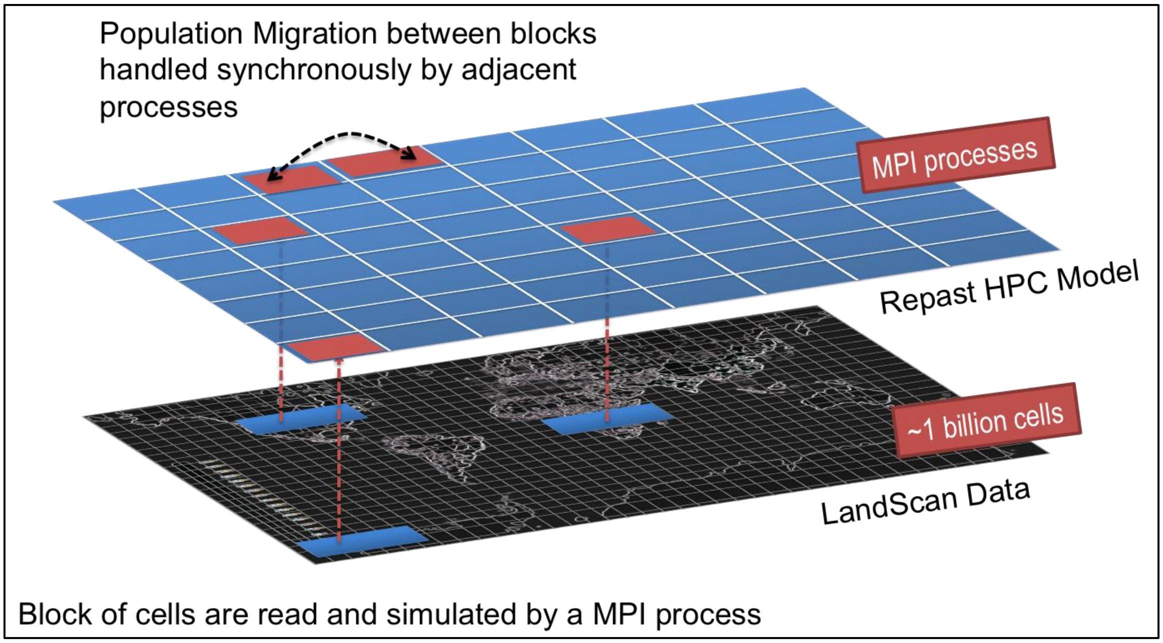

Population Migration between blocks handled synchronously by adjacent processes
MPI processes
Repast HPC Model
~1 billion cells
LandScan Data
Block of cells are read and simulated by a MPI process

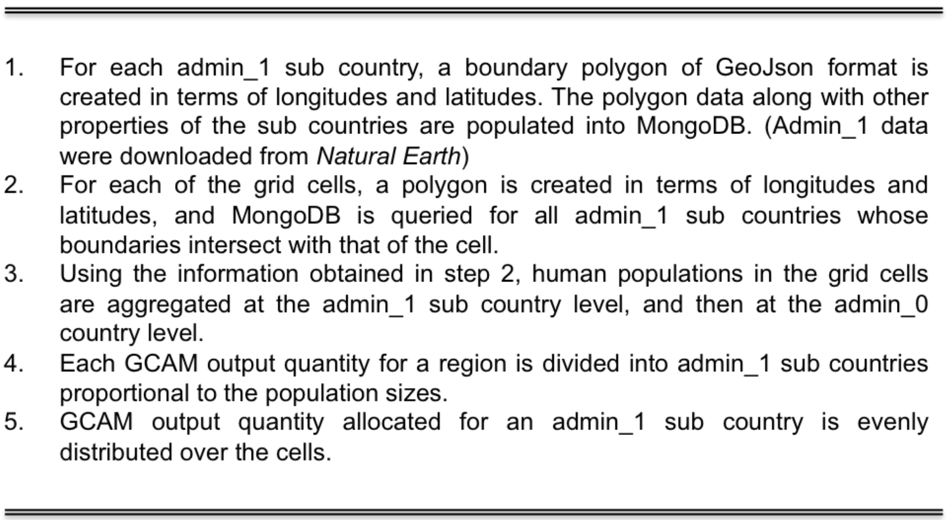
Figure 2. Repast HPC with LandScan Data

GCAM (Global Change Assessment Model), Version 4, is a dynamic recursive serial model, written in C++ with continuing development by Pacific Northwest National Laboratory, which includes as drivers representations of the global economy, land use, agriculture, energy system, and climate. In GCAM, potential gross domestic product is computed based on labor productivity and population size in each of 32 regions spanning the globe. GCAM integrates energy, terrestrial carbon cycle, agriculture, land market, and forestry computations across 151 agro-ecological zones to calculate simultaneous market-clearing prices under predetermined policy scenarios at five-year time intervals (Bond-Lamberty et al., 2014; Voisin et al., 2013). For our experiment, we extracted the corn production from the model at the 32-region resolution.

The Repast model is implemented using the Repast for High Performance Computing (Repast HPC) library. Repast HPC is a free and open source agent-based modeling and simulation (ABMS) library developed at Argonne National Laboratory for high performance distributed computing platforms (http://repast.sourceforge.net). Repast HPC is written in C++ using the Message Passing Interface (MPI).

In the MIRAGE framework, the Repast model reads LandScan data and places the human population in appropriate grid cells. As shown in fig. 2, an MPI-process of the Repast model is in charge of simulating actors in a block of cells. MPI processes collectively handle the migrations of populations across adjacent blocks.

The Repast model extends a pedagogical and demonstrative Repast HPC model in which susceptible and infected humans are placed in a grid space and the former try to avoid the latter. The simulation progresses as the global clock advances forward. At each simulated clock tick, susceptible humans move to one of the neighboring grid cells where infected humans are least common, and the infected humans move to where the susceptible humans are most common. When the two groups come into direct contact, infection takes place. Infection takes place within a short interval. This basic model has been extended to incorporate additional data into the simulation so that more elaborate scenarios can be modeled. In the demonstration presented here, corn is introduced as a resource for susceptible humans, and the resulting dynamics are generated through infected/susceptible/corn interactions.



1. For each admin_1 sub country, a boundary polygon of GeoJson format is created in terms of longitudes and latitudes. The polygon data along with other properties of the sub countries are populated into MongoDB. (Admin_1 data were downloaded from *Natural Earth*)
2. For each of the grid cells, a polygon is created in terms of longitudes and latitudes, and MongoDB is queried for all admin_1 sub countries whose boundaries intersect with that of the cell.
3. Using the information obtained in step 2, human populations in the grid cells are aggregated at the admin_1 sub country level, and then at the admin_0 country level.
4. Each GCAM output quantity for a region is divided into admin_1 sub countries proportional to the population sizes.
5. GCAM output quantity allocated for an admin_1 sub country is evenly distributed over the cells.

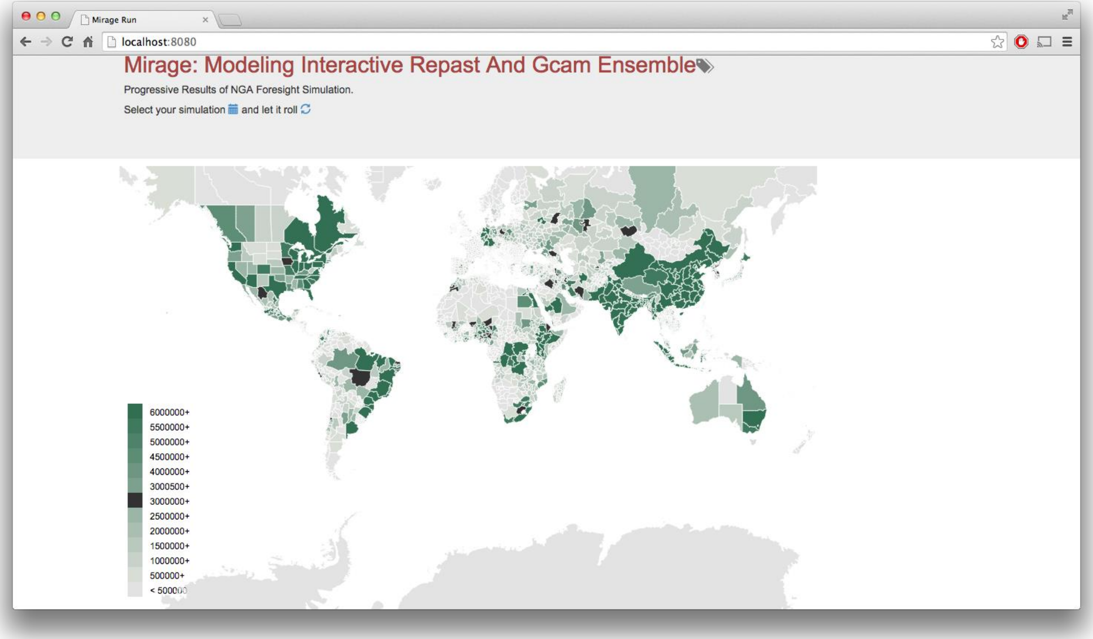
Table 1. Procedures to map data between GCAM and the Repast model.

## 3. Model execution and the work in progress

The initial inflow of GCAM data (i.e., corn yields) feeds into the Repast model that in turn simulates agent interaction dynamics and produces outputs after finishing a predefined number of iterations. Currently, the reading of infected populations and corn yields, as well as susceptible populations, are done through files. For consistency and efficiency, we use the hierarchical data format version 5 (HDF5) as the format for all data files and represent all data as numbers in grid cells stored as HDF5 matrices. Each MPI process reads the input for only its portion of the data using collective MPI-IO.

With resolution of 1 km, LandScan places the initial susceptible population of approximately 7 billion over a 21,600 x 43,200 grid of cells. Running the Repast model at the full resolution of LandScan is challenging even for leading HPC facilities. Thus, while we are planning full-scale runs on the Titan supercomputer of the Oak Ridge Leadership Computing Facility, we evaluated the model at a reduced resolution. More specifically, LandScan resolution was reduced to 360 x 720 cells, each of which spans half degree in both longitude and latitude. Since GCAM produces outputs for 32 regions of the earth, it is required that GCAM outputs be divided over the 360 x 720 grid cells. Table 1 describes the allocation process.

To facilitate understanding of a simulation and its progression over time, a visual feedback system was designed. The system uses a web-based interface through which users can retrieve their simulation results and watch the progression of parameters of interest over time. Currently, susceptible/infected population changes, at the sub-country (i.e., administrative level one, or “admin_1”) level, can be monitored and visualized at the same resolution. For this, the Repast model produces snapshots of population of each grid cell at a regular interval. These grid-based data are mapped to admin_1 level following a similar procedure above and stored in the MongoDB server. The visualization front-end is implemented using D3 and Node.js. fig. 3 shows a screen shot of the visual feedback.



Mirage: Modeling Interactive Repast And Gcam Ensemble

Progressive Results of NGA Foresight Simulation.

Select your simulation and let it roll

Figure 3. Visual Feedback from the MIRAGE Framework.

## 4. Conclusion and the next steps

This project represents a first pass at integrating agent-based modeling, global change assessment modeling, and geospatial data with an eye toward evaluating global change scenarios and their possible impacts. We have shown that a reasonable framework can be built that allows us to combine serial and parallel models that output different file types and incorporates data at different spatial scales to produce a result that provides new information given an established initial parameter set. The model will include two-way coupling and the potential to exchange, in a scientifically robust way, additional parameters among the currently coupled and planned models.

## 5. Acknowledgements

This manuscript has been authored in part by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. This manuscript has also been authored in part by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated under contract number DE-AC02-06CH11357. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide

license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

## 6. References

Bond-Lamberty B, Calvin K, Jones A, Mao J, Patel P, Shi X, Thomson A, Thornton P, Zhou Y, 2014, Coupling earth system and integrated assessment models: the problem of steady state. *Geoscientific Model Development*, 7, 1499-1524.

Collier N, and North M, 2012, Parallel agent-based simulation with Repast for High Performance Computing. *Simulation*, 0(0):1-21.

McKee J, Rose A, Bright E, and Huynh T, 2015, A Locally-Adaptive, Spatially-Explicit Projection of U.S. Population for 2030 and 2050. *Proceedings of the National Academy of Sciences*, In Press.

Voisin N, Liu L, Hejazi M, Tesfa T, Li H, Huang M, Liu Y, and Leung L, 2013, One-way coupling of an integrated assessment model and a water resources model: evaluation and implications of future changes over the US Midwest. *Hydrology and Earth System Sciences*, 17, 4555-4575.