G·E·O·S·H·A·R·E

An introduction to GEOSHARE's cyberinfrastructure: exchange and analysis of geo-spatial data through HUBzero

Nelson Villoria, GEOSHARE and NSF-DIBBS Co-Principal Investigator

Department of Agricultural Economics

Purdue University

GLP 2nd Open Science Meeting, Berlin March 21, 2014

This presentation is based on work with the following collaborators:

- HongJun Choi (Purdue U.)
- Delphine Derying (University of East Anglia)
- Joshua Elliot (AgMip)
- Thomas Hertel (Purdue U.)
- Erich Huebner (Purdue U.)
- Rajesh Kalyanam (Purdue U.)
- Danny Plouffe (McGill U.)
- Navin Ramankutty (McGill U.)
- Carol Song (Purdue U.)
- Lan Zhao (Purdue U.)

GEOSHARE's Mission

GEOSHARE is a decentralized network of scientists in many disciplines which mission is to provide:

- a freely available, global, spatially explicit database on agriculture, land use, and the environment
- analysis tools for scientists, decision makers, and development practitioners

GEOSHARE: Vision of a Network

GEOSHARE envisions a vibrant global network:

- contributing to shared cyberinfrastructure
- enhancing capacity for geospatial analysis
- applying geospatial tools to guide decision making related to food security, land use, environmental sustainability and poverty reduction

GEOSHARE's Network Today (Pilot Project)

Global Land Use
Navin Ramankutty
McGill U.

Global Water Use in Agriculture
Stefan Siebert, U. Bonn

Global Climate
Noah Diffenbaugh
Stanford U.

GEOSHARE
Economic Analysis and Project Coordination
Thomas Hertel and Nelson Villoria
Purdue University

Agriculture, Environment, & Poverty in Asia
Andrew Nelson, IRRI

Agriculture, Environment, & Poverty in Latin America
Glenn Hyman, CIAT

Agriculture, Environment, & Poverty in Africa
Jawoo Koo, IFPRI

Cyberinfrastructure for GEOSPATIAL Analysis
Carol Song, Purdue University

Why a shared cyberinfrastructure for GEOSHARE?

Different Worlds

Expert groups in educational and international institutions generate knowledge, data, and tools

Often, knowledge stays within disciplinary silos

Crop Models

Users elsewhere wish to use knowledge for decision making, learning, etc.

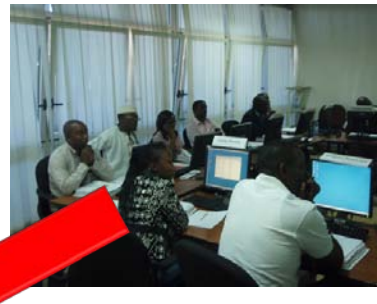
What Keeps Them Separate?

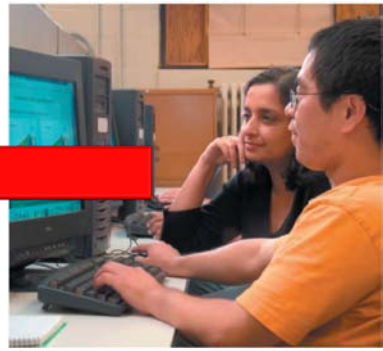
Commercial Software
is expensive, requires powerful
Hardware, and specialization
SAS ≠ STATA ≠ ARCGis

Multidisciplinary
collaborations to
open the silos are
costly

HUBzero technology seeks to break these barriers

Use the Web to share data, code, hardware, and spread the impact of research efforts and funds





No additional Software to install or learn

How to do this?

- hubzero PLATFORM FOR SCIENTIFIC COLLABORATION
- HUBzero is a technology developed at Purdue University with NSF funds for creating dynamic websites for scientific research and educational activities:
  - It's fully open source
  - Whatever can be compiled in Linux, can be run in the Hub
  - Rapid deployment of Graphical User Interfaces
  - Access to cluster computer resources

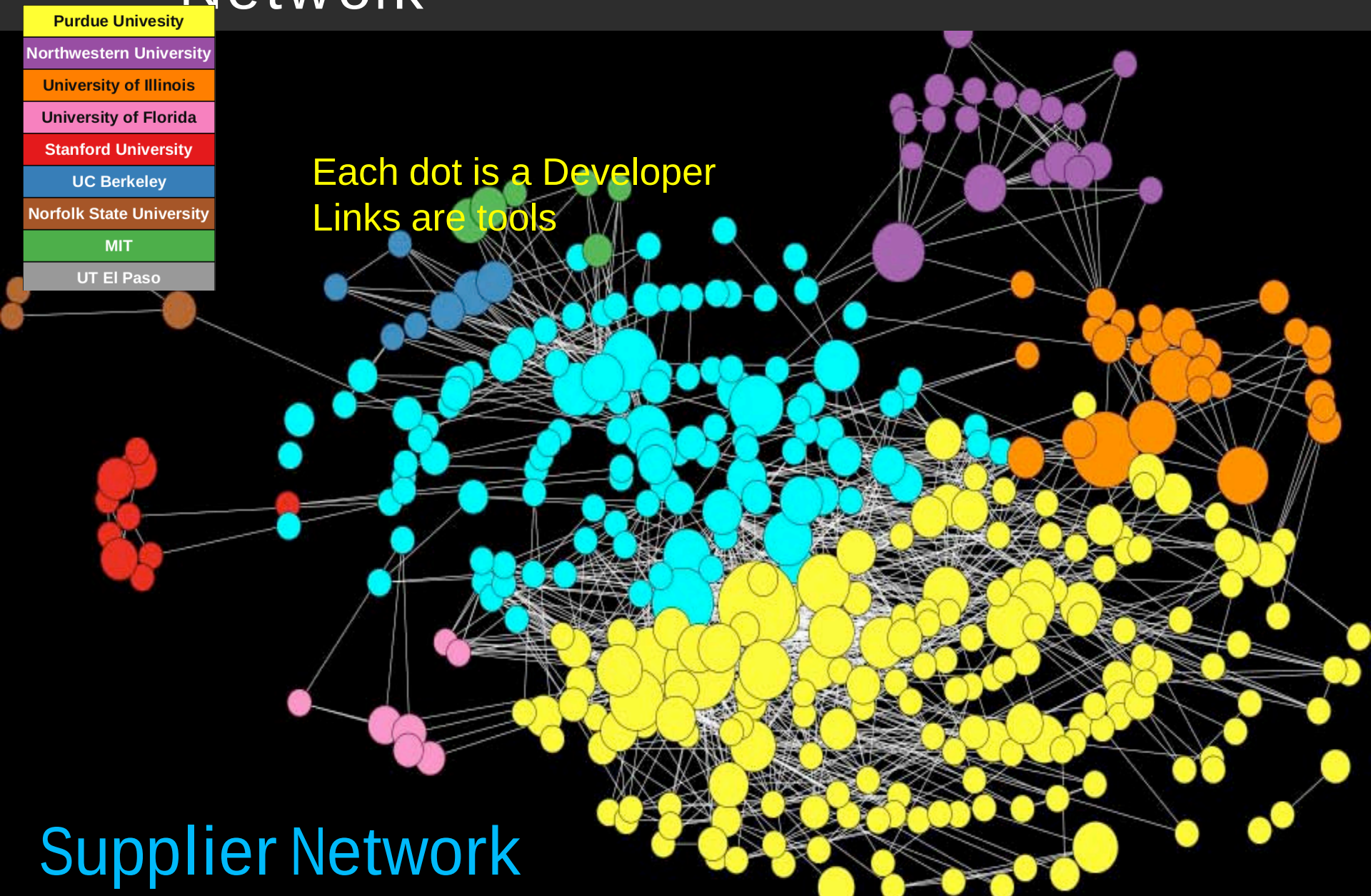
Have this worked elsewhere?

- We have been inspired by nanoHUB.org
- nanoHUB.org's objective is to transfer the knowledge generated in extensive nanotechnology facilities to others in academy and industry
- Created by the NSF-funded Network for Computational Nanotechnology.
- Next slides from Prof. Gerhard Klimeck, nanoHUB.org Director

The nanoHUB Proposition and Emerged (and Busted) Myths

Proposition	Myth:
Be Developer Friendly	Cannot use research codes for education Must write own code to do research Experimentalists cannot use research codes
Be Accessible	NO End-to-end Science Cloud Possible
Be User Friendly	Building User Interfaces too Difficult Must rewrite code for web deployment There is no incentive to share codes

nanoHUB.org: Developer Collaboration Network



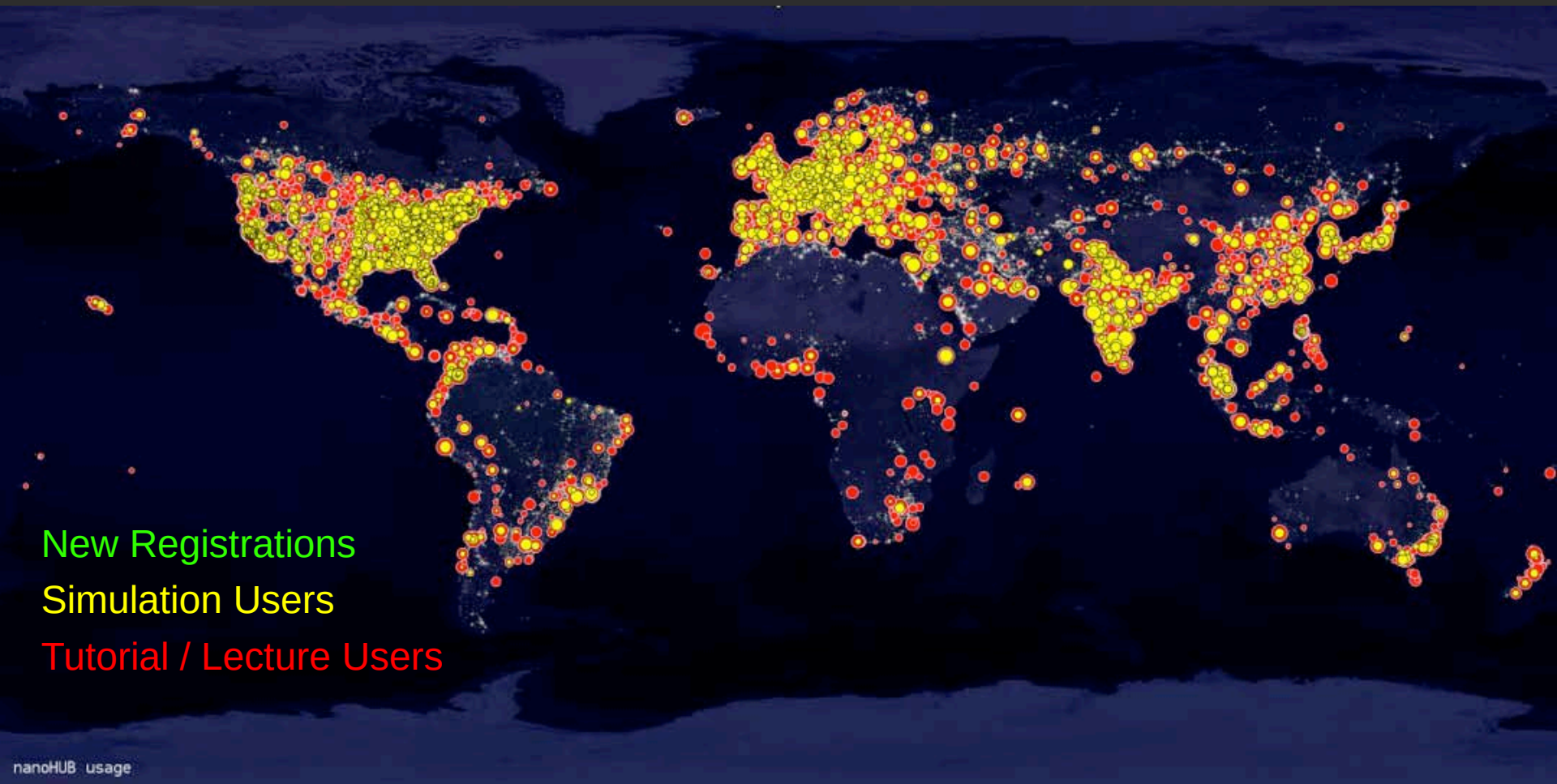
Supplier Network

http://nanoHUB.org

Network members in 172 countries

Over 12,000 simulations/year

Over 300,000 users annually



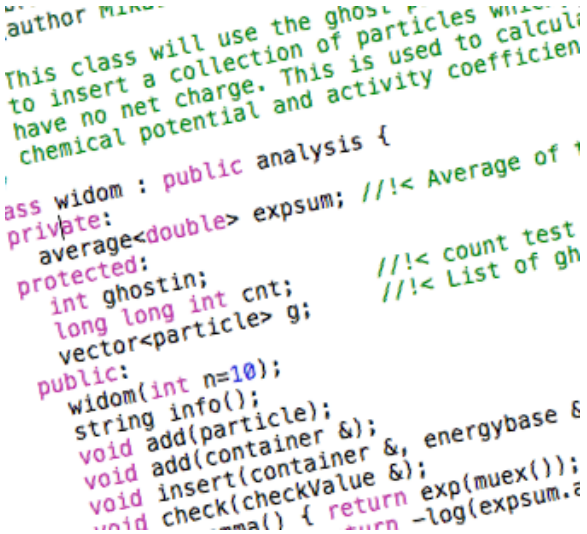
GEOSHARE Example 1: Pegasus

Pegasus 1.0 Predicting Ecosystem Goods And Services Using Scenarios

- Published by/in:
  - Deryng, D., W. J. Sacks, C. C. Barford, and N. Ramankutty. 2011. “Simulating the Effects of Climate and Agricultural Management Practices on Global Crop Yield.” *Global Biogeochemical Cycles* 25 (May): 18 PP.
- A global crop model that integrates climate, the effect of planting dates and cultivar choices, irrigation, and fertilizer application on crop yield for maize, soybean, and spring wheat.
- Useful for studying adaptation to climate change.

Inputs when we started:

- Fortran Code with Model Equations:



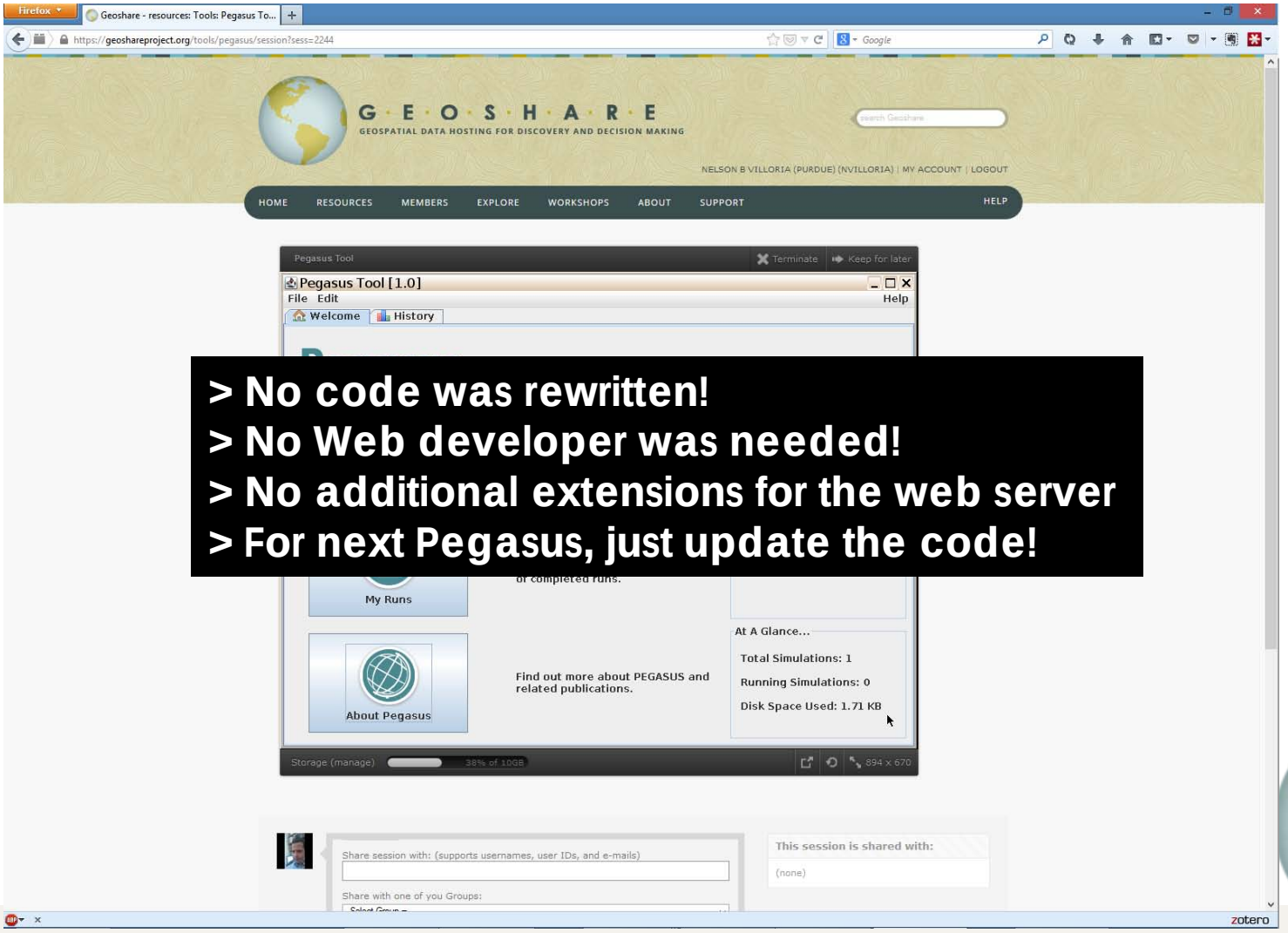
```
author MIKE

This class will use the ghost
to insert a collection of particles
have no net charge. This is used to calcul
chemical potential and activity coefficien

ass widom : public analysis {
private:
  average<double> expsum; //!< Average of
protected:
  int ghostin;
  long long int cnt;          //!< count test
  vector<particle> g;         //!< List of gh
public:
  widom(int n=10);
  string info();
  void add(particle);
  void add(container &);
  void insert(container &, energybase &
  void check(checkValue &);
  mma() { return exp(muex());
  return -log(expsum.a
```

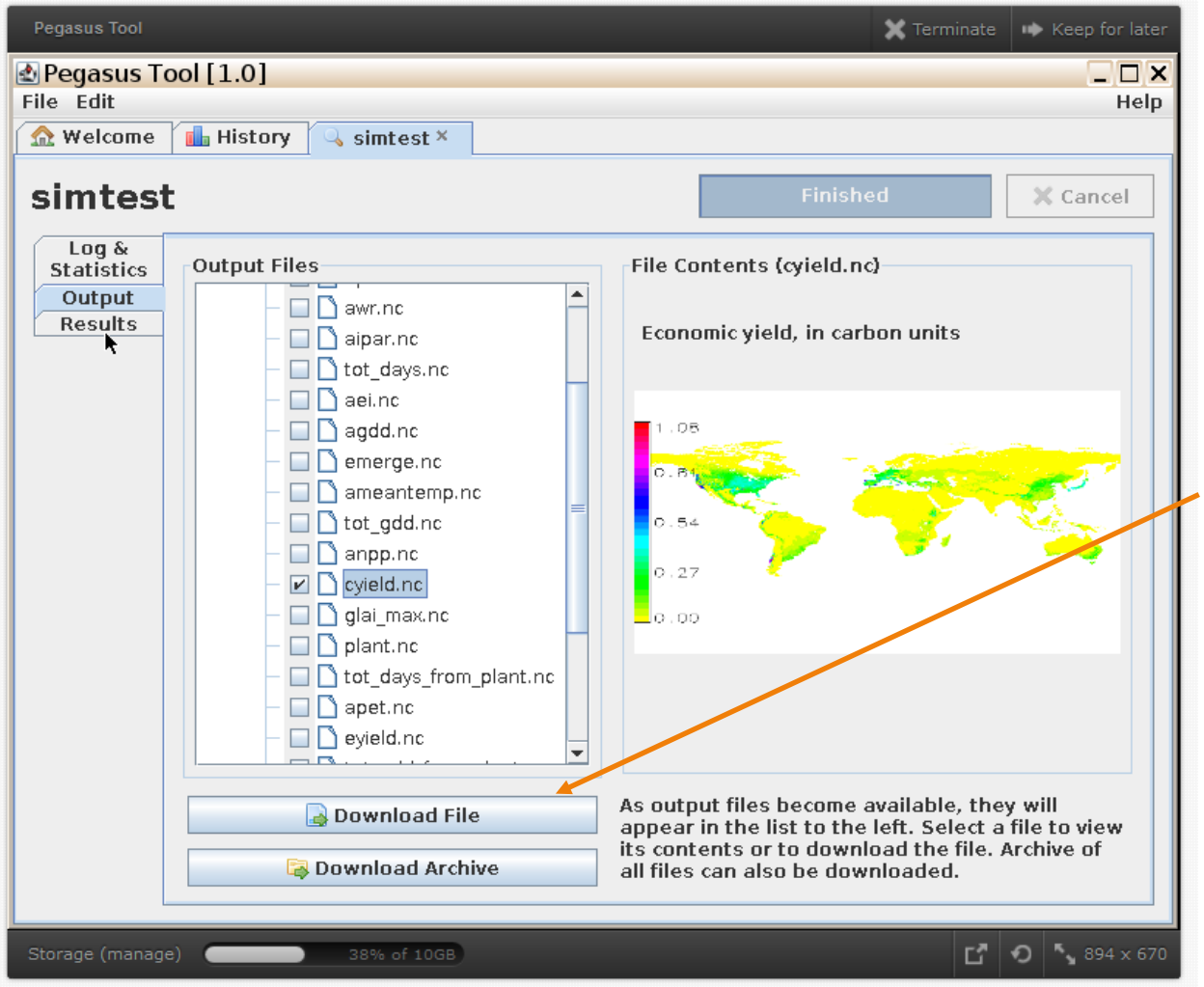
- Few megabytes of data in NetCDF format to calibrate the model.

Now: Pegasus 1.0 at geoshareproject.org



> No code was rewritten!
> No Web developer was needed!
> No additional extensions for the web server
> For next Pegasus, just update the code!

On-the-fly visualization of results



Download output for Further processing

GEOSHARE Example 2: Processing AgMIP's Aggregated Data

The AgMIP GRIDded crop modeling initiative

- Crop Modeling Teams grouped around AgMIP The Agricultural Model Intercomparison and Improvement Project
- The Global Gridded Crop Model Intercomparison (GGCMI):
  - Fast track phase:
    - 5 Crop Models
    - 5 Global General Circulation Models
    - 5 Representative Concentration Pathways
    - CO2 Fertilization and irrigation
    - 12 Crops
    - All the combinations ran from 1971 to 2099
- + 35,000 0.5^0x0.5^0 grids with crop yields

The Problem

- Economic modelers use data aggregated to the country level
- NetCDF and other spatial data delivery formats are foreign to non-specialists
- Download of large grids is challenging even with good bandwith
- Storage is challenging for otherwise very good PCs
- How to make the data available to Users

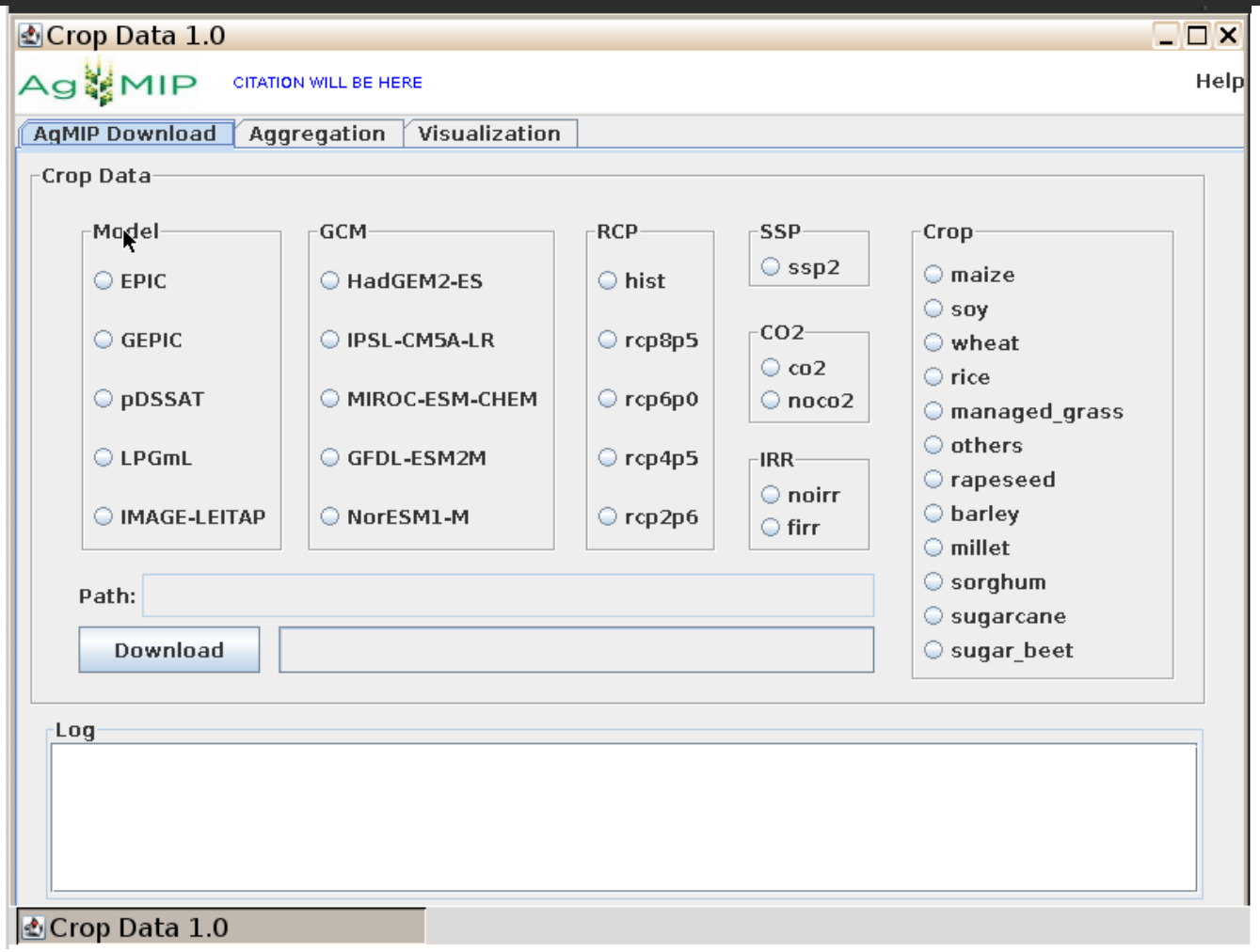
The Solution

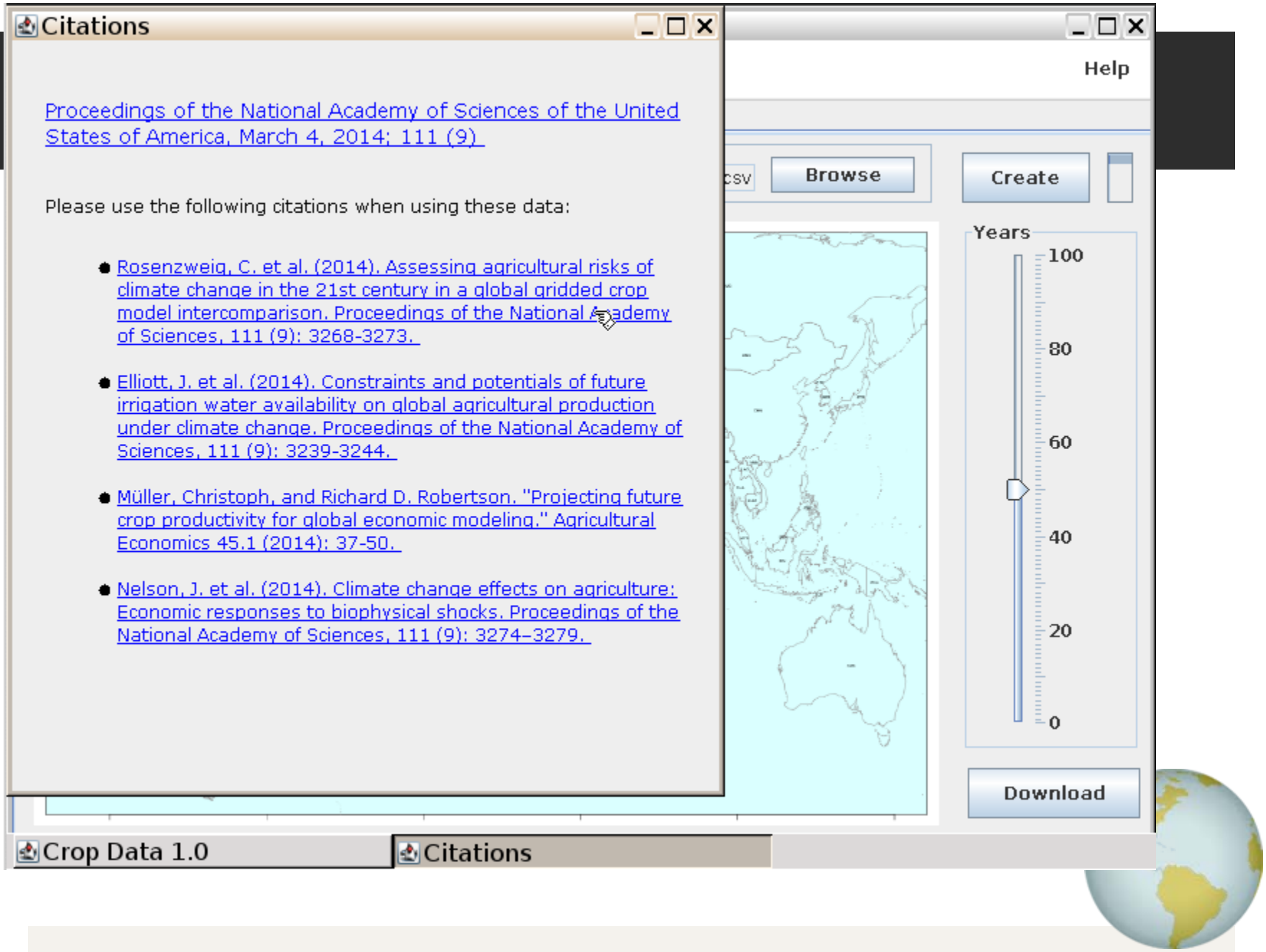
- Good all-around R Code:

  - Reads data from Globus Online to the HUB

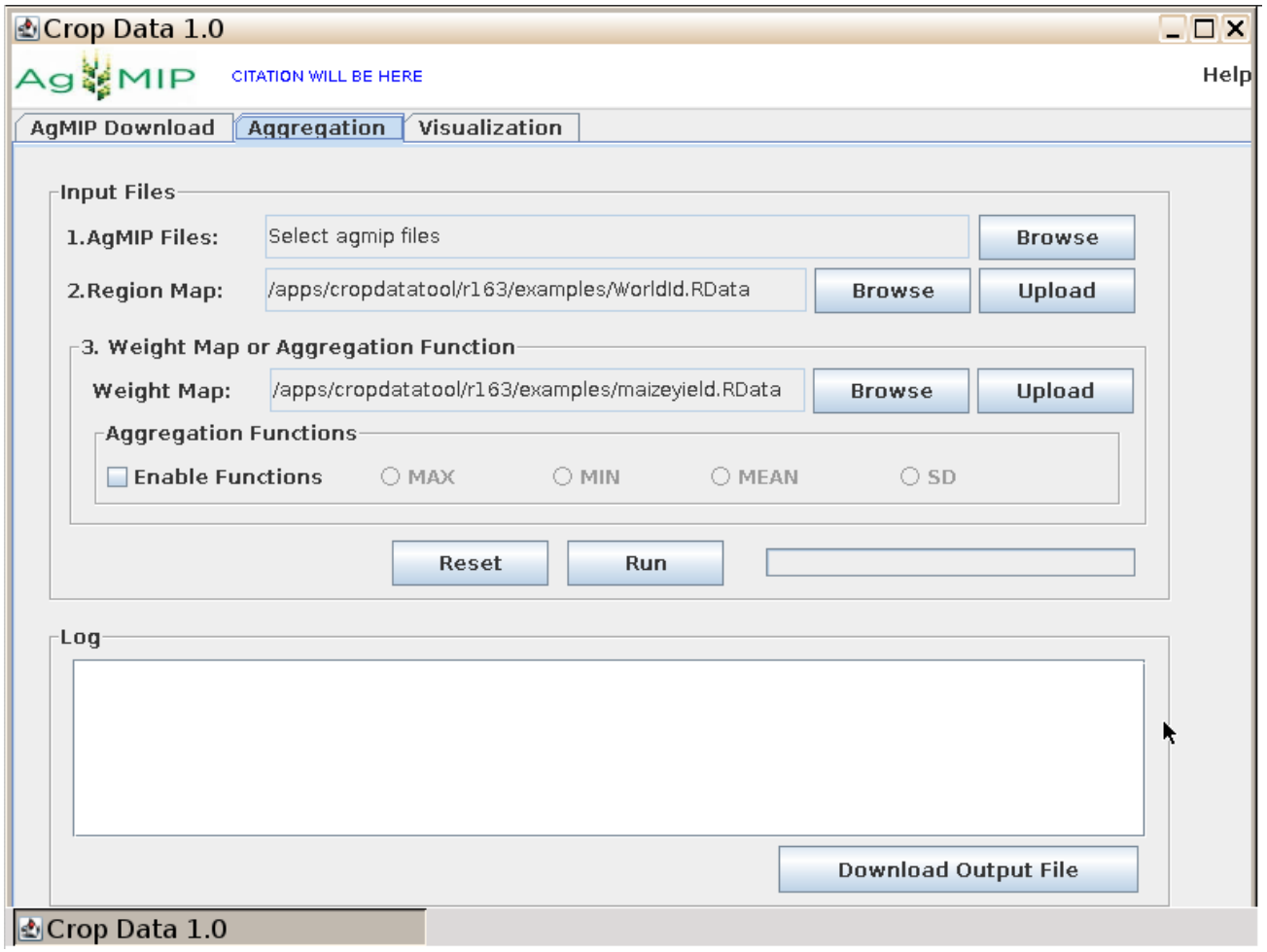
  - Use aggregate() function
  - Aggregates using mapping from XY coordinate to country, AEZs, up to the user.
  - Calculate summary statistics (mean, etc.) or weighted averages using user provided weights (production, area, population, etc.)
  - Wrap the R function around a Graphical User interface

The Result

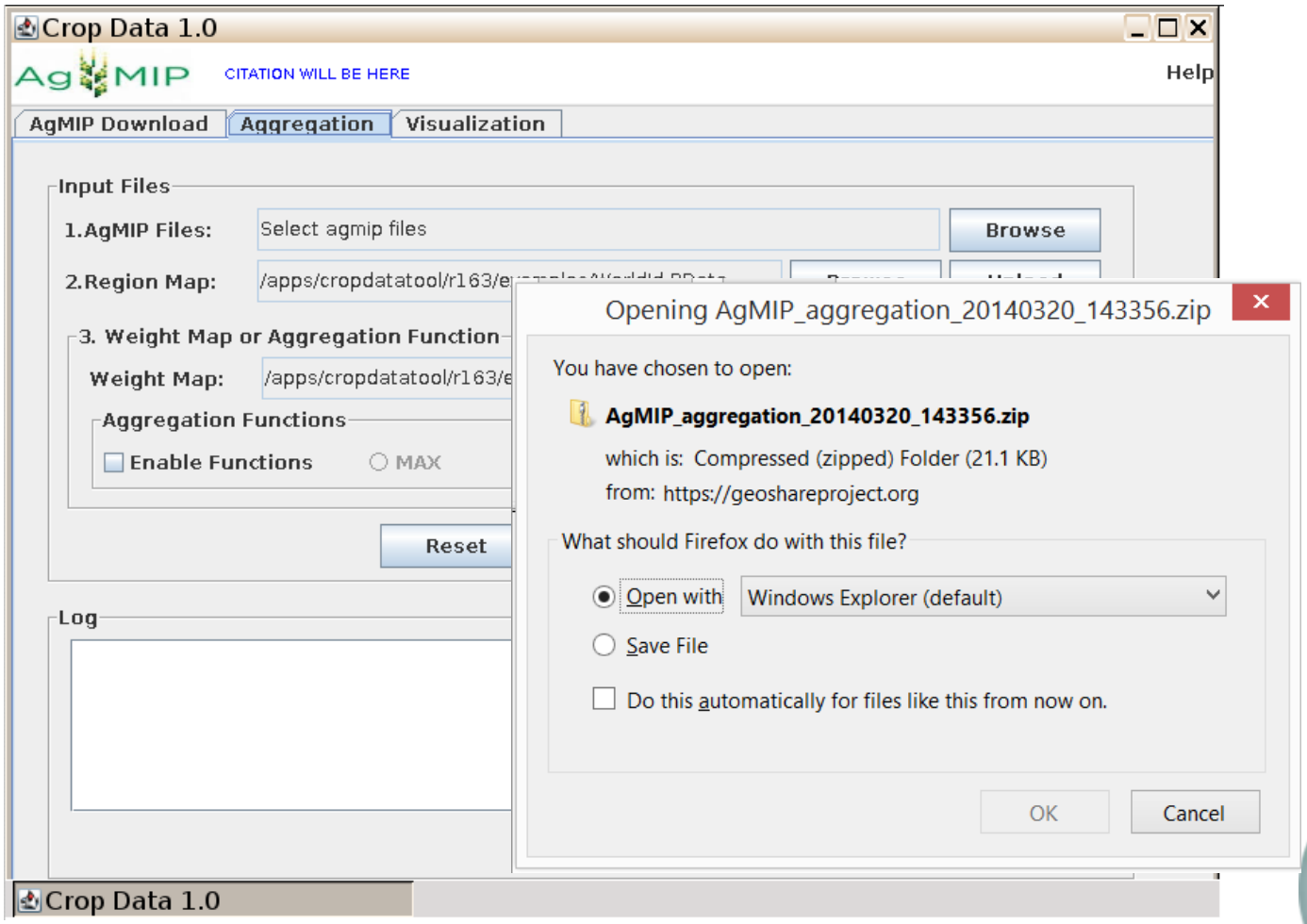

Crop Data 1.0
AgMIP
CITATION WILL BE HERE
Help
AgMIP Download
Aggregation
Visualization
Crop Data
Model
EPIC
GEPIC
pDSSAT
LPGmL
IMAGE-LEITAP
GCM
HadGEM2-ES
IPSL-CM5A-LR
MIROC-ESM-CHEM
GFDL-ESM2M
NorESM1-M
RCP
hist
rcp8p5
rcp6p0
rcp4p5
rcp2p6
SSP
ssp2
CO2
co2
noco2
IRR
noirr
firr
Crop
maize
soy
wheat
rice
managed_grass
others
rapeseed
barley
millet
sorghum
sugarcane
sugar_beet
Path:
Download
Log
Crop Data 1.0


Citations
Proceedings of the National Academy of Sciences of the United States of America, March 4, 2014; 111 (9)
Please use the following citations when using these data:
Rosenzweig, C. et al. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proceedings of the National Academy of Sciences, 111 (9): 3268-3273.
Elliott, J. et al. (2014). Constraints and potentials of future irrigation water availability on global agricultural production under climate change. Proceedings of the National Academy of Sciences, 111 (9): 3239-3244.
Müller, Christoph, and Richard D. Robertson. "Projecting future crop productivity for global economic modeling." Agricultural Economics 45.1 (2014): 37-50.
Nelson, J. et al. (2014). Climate change effects on agriculture: Economic responses to biophysical shocks. Proceedings of the National Academy of Sciences, 111 (9): 3274–3279.
Help
Browse
Create
Years
100
80
60
40
20
0
Download
Crop Data 1.0
Citations

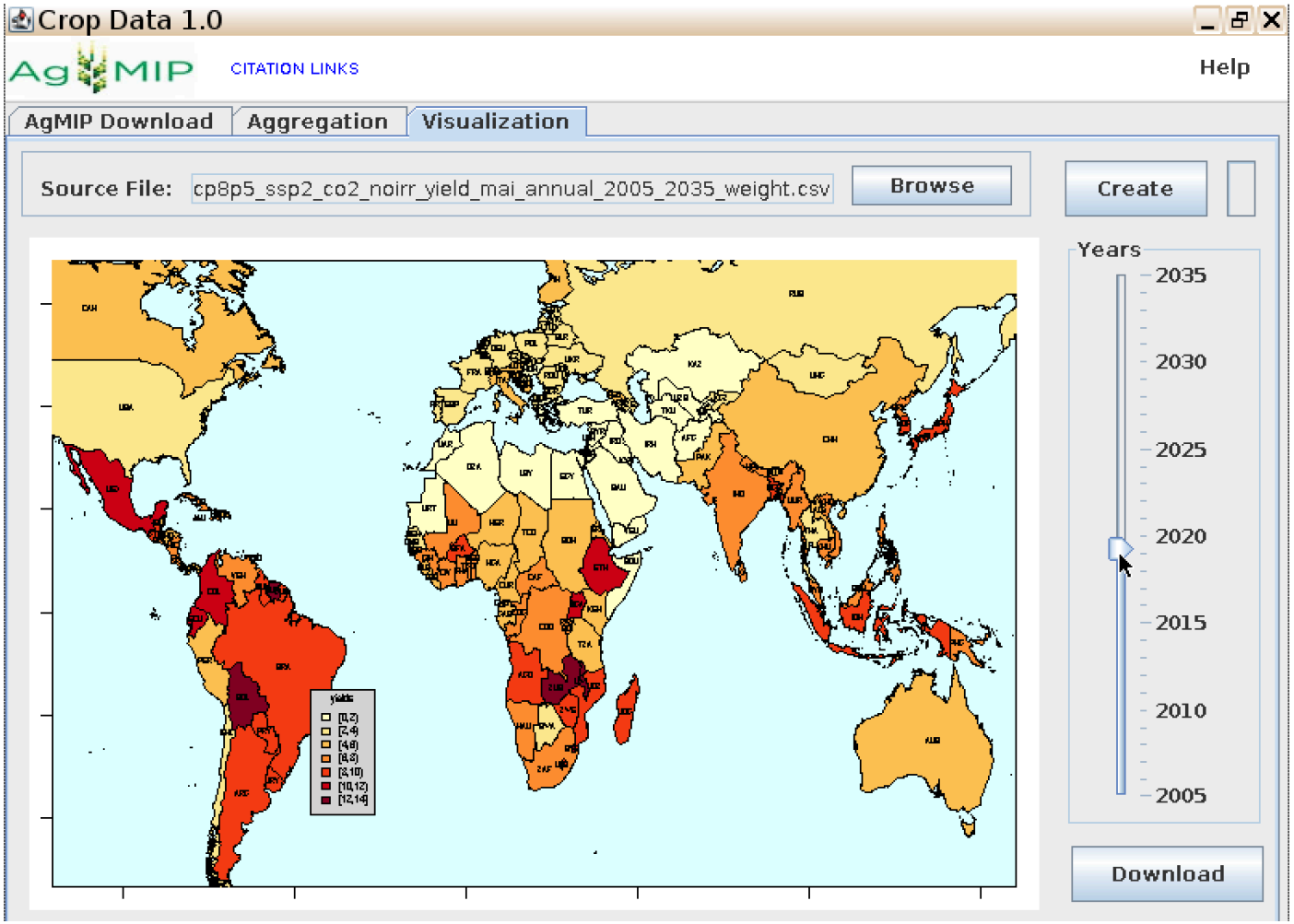
Aggregation Possibilities



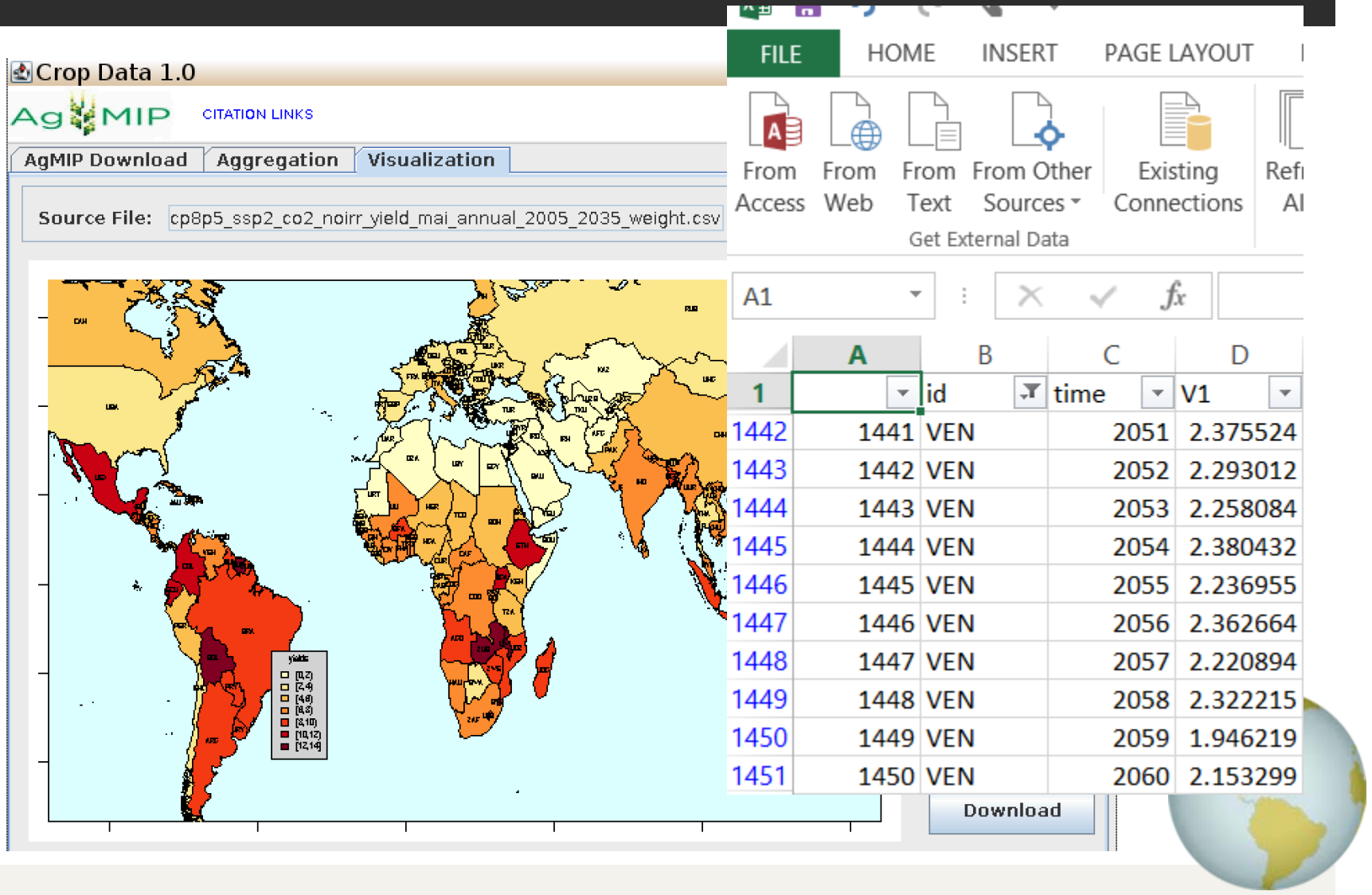
Aggregation Possibilities



Visualization



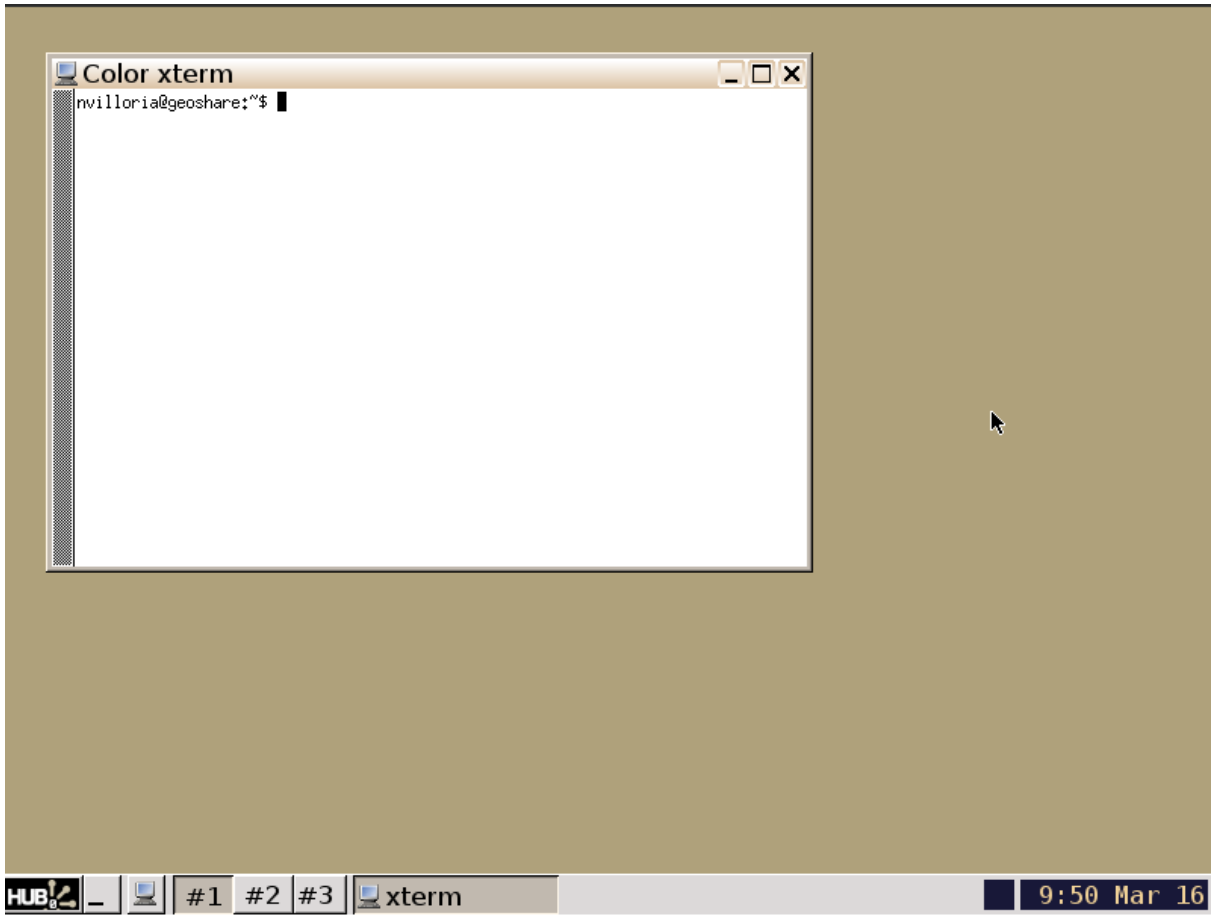
Visualization



Costs and Impacts

- Costs to the developers:
  - Write R function
  - Test it
  - Develop GUI
- Cost to the User:
  - None.
- Software Licenses, Hardware and Other Equipment
  - None.
  - All the processing on-the-fly and on university owned hardware
- Impacts:
  - Anyone interested in getting aggregated climate shocks for aggregated analysis can do so.

A peak inside the Hub: A Linux Workspace

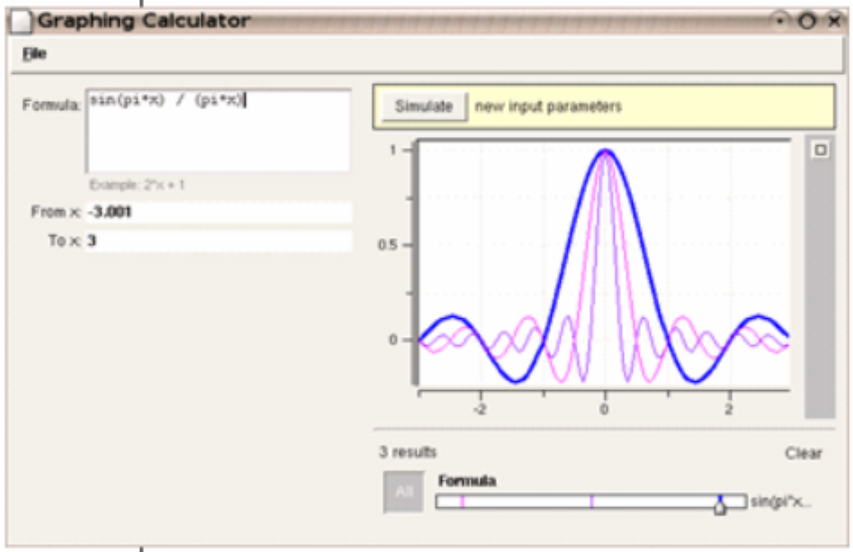


If it can be run on Linux, it can be converted on a web application

Graphical User Interfaces

- HUBzero supports its own language, XML based, called RAPPTURE toolkit
- But also supports Python, Java, etc.
- User's knowledge and experience set boundaries.

```
<?xml version="1.0"?>
<run>
  <tool>
    <title>Graphing Calculator</title>
    <about>Press Simulate to view results.</about>
    <command>python @tool/graph.py @driver</command>
  </tool>
  <input>
    <string id="formula">
      <about>
        <label>Formula</label>
        <hints>Example: 2*x + 1</hints>
      </about>
      <size>30x5</size>
    </string>
    <number id="min">
      <about> <label>From x</label> </about>
      <default>0</default>
    </number>
    <number id="max">
      <about> <label>To x</label> </about>
      <default>1</default>
    </number>
  </input>
  <output>
    <curve id="result">
      <about> <label>Formula: Y vs X</label> </about>
    </curve>
  </output>
</run>
```

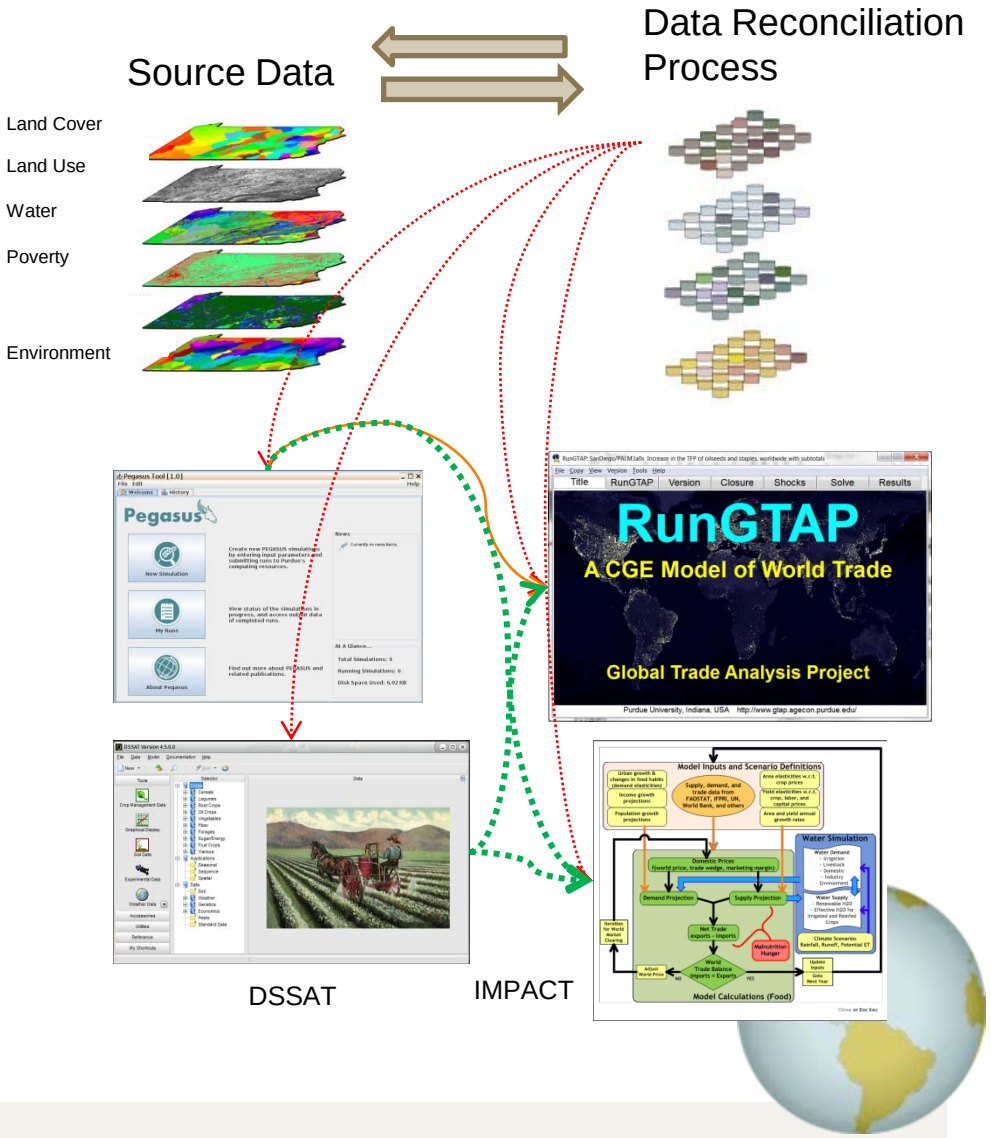


GEOSHARE Example 3: A workflow for assessing the value of improved data

Workflows: From data to analysis

- Gridded source data on land cover & use, water, poverty and environment - *flows into* -
- Data reconciliation 'models', e.g. SPAM which produce usable data - *for use in* -
- Biophysical and economic models (e.g., DSSAT, IMPACT, GTAP)

  Sensitivity of results at final stage determines value of improving quality of source data



Other considerations

GEOSHARE is free for all users and open to all

- Yes, users pay nothing.
- What about developers?
  - This is shared space
  - The cost of the Hub is around 50,000 per/year
  - Working on GEOSHARE governance cost-sharing model for self sustainability
  - But opportunities for economies of scale are huge.

Recent Funding For Expanding HUBzero's Spatial Capabilities

- A National Science Foundation grant
- Data Infrastructure Building Blocks (DIBBs) program
- GABBs: 1 of 4 implementation awards in 2013
- $4.5M over 4 years
- Started October 1, 2013
- Collaboration with other awards

Sister geospatial hub projects

- Efforts in developing integrated geospatial data/modeling capabilities using HUBzero
  - Drinet hub (http://drinet.hubzero.org)
  - Geoshare hub (http://geoshareproject.org)
  - Water hub (http://water-hub.org)
  - Useful to Useable (u2u) http://drinet.hubzero.org/groups/u2u

Overarching goals

- Enabling geospatial modeling and analysis online
- Anyone can create an online geospatial app and share
- Further support for geospatial data
- Building blocks can be used by other projects

Building for self service (DIY) – Leverage successful software – Develop building blocks

- Questions
- Feedback?

- Thanks!

- nvillori@purdue.edu