Rapid aggregation of global gridded crop model outputs to facilitate cross-disciplinary analysis of climate change impacts in agriculture

Nelson B. Villoria[a,*], Joshua Elliott[b], Christoph Müller[c], Jaewoo Shin[d], Lan Zhao[e], Carol Song[e]

[a]*Department of Agricultural Economics, Purdue University, 403 West State Street, West Lafayette, IN 47907, USA.*
[b]*University of Chicago and ANL Computation Institute Center for Robust Decision-making in Climate and Energy Policy (RDCEP) 5735 S. Ellis Ave. Chicago, IL 60637, USA.*
[c]*Potsdam Institute for Climate Impact Research, Telegraphenberg A 31 14473, Potsdam, Germany.*
[d]*Department of Research Computing, Purdue University, West Lafayette, IN 47907, USA.*
[e]*Rosen Center for Advanced Computing, Purdue University, West Lafayette, IN 47907, USA.*

Keywords: Crop Models, AgMIP, GEOSHARE, Global Agriculture, Climate Change

Abstract

We discuss an on-line tool that facilitates access to the large collection of climate impacts on crop yields produced by the Agricultural Model Intercomparison and Improvement Project. This collection comprises the output of seven crop models which were run on a global grid using climate data from five different general circulation models (GCM) under the current set of representative pathways (RCP). The output of this modeling endeavor consists of more than 36,000 publicly available global grids at a spatial resolution of one half degree. The tool is implemented in GEOSHARE's HUBzero cyberinfrastructure. We offer flexible ways to aggregate these data while reducing the technical barriers implied by learning new download platforms and specialized formats. The tool is accessed trough any standard web browser without any special bandwidth requirement.

Software Availability

AgMIP @ GEOSHARE: A GEOSHARE tool for aggregating outputs from the AgMIP's Global Gridded Crop Model Intercomparison Project, first released in March 2014, is freely available at the GEOSHARE website (`https://mygeohub.org/groups/geoshare/resources?area=tools`) using any standard Internet browser. All the programs—a java graphical user interface (GUI) and a set of R functions— can be freely downloaded and reused. The tool is free under a GNU General Public License (www.gnu.org) agreement. Documentation and support for users include a User's Manual, as well as a set of default regional maps and weighting schemes.

---

[*] Corresponding author. Department of Agricultural Economics Purdue University 403 West State Street, Krannert Bldg. West Lafayette, IN 47907, USA. Email: nvillori@purdue.edu.

1. Introduction

We discuss an online tool that facilitates access to a large collection of climate impacts on crop yields produced by the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al., 2013) as part of the Global Gridded Crop Model Intercomparison Initiative (GGCMI; Elliott et al., 2014c) and the Inter-Sectoral Impacts Model Intercomparison Project (ISI-MIP; Warszawski et al., 2014). The reader is advised to consult Rosenzweig et al. (2014) for an introduction to these data and for further references. As displayed in table 1, this collection comprises the output of seven crop models which were run on a global grid using climate data from five different general circulation models (GCM) under the current set of representative pathways (RCP). The output of this modeling endeavor consists of more than 36,000 publicly available global grids at a spatial resolution of one half degree.

Table 1: Models and crops

Model	Crops
EPIC	All
GEPIC	wheat, maize, soy, rice
pDSSAT	wheat, maize, soy
LPJmL	All
IMAGE-AEZ	All
PEGASUS	wheat, maize, soy
LPJ-GUESS	wheat, maize, soy

Notes. All crops: maize, soybeans, wheat, rice, managed grass, rapeseed, barley, millet, sorghum, sugarcane, sugar beets and others. In addition, each crop model was run under five different scenarios (historical and four representative concentration pathways) generated by the global circulation models HadGEM2-Es, IPSL-CM5A-LR, MIROC-ESM-CHEM, GFDL-ESM2M, NorESM1-M.

This information has been used to gain an understanding of the interactions among water supply, irrigation, and climate change in global caloric production (Elliott et al., 2014a); multisectoral impacts of climate change (Piontek et al., 2014); and endogenous economic responses to increases in temperature (Nelson et al., 2014). The wide range of applications of these data can be expected to greatly expand the quantitative assessment of global climate change impacts at different levels of global warming as well as geographic scales. Importantly, the richness in crops and GCM models can inform our understanding of model uncertainty (e.g., Piontek et al., 2014) and therefore help to impose boundaries on the potential effects of climate warming as well as to illuminate research priorities. However, and despite being open access, the technical skills required to access these data are likely to represent an important barrier for many researchers, reducing the potential impact of this information.

For many applications, the relevant unit of observation may be several grid-cells within a political administrative unit such as a country, perhaps combined with agroecological zones (Monfreda et al., 2009); therefore, a crucial step for utilizing these data is meaningful aggregation across temporal and spatial scales. However, aggregation of data with high spatial and temporary resolution requires specialized knowledge about geoprocessing procedures (Hofer, 2014) as well as domain-specific definitions of relevant spatial weights (Stasch et al., 2014). Moreover, data acquisition—searching, downloading, and storing and maintaining these data—requires costly hardware, costly software licenses, good Internet connectivity, and specialized staff (Hertel et al., 2010).

A typical user of the GGCM archive would need to set up a Globus Online (Foster, 2011) client, search the desired dataset in a multi-layered folder hierarchy, download the data, and use specialized tools to extract the information from the NetCDF files in which the data are stored. Given the potentially large volume of information, data download and storaging may consume significant bandwidth and hardware resources. Aggregation from the grid-cells to the desired geographic units requires significant dexterity along with specialized geoprocessing tools that involve constructing aggregation weights as well as concordances between coordinates and the desired geographic regions.

Of course, these difficulties are not exclusive to the GGCM archive. For instance, Hertel et al. (2010) point out that technical access barriers have slowed down our understanding of the effects of global environmental change on the long-run sustainability of the food system. From a more general perspective, Craglia et al. (2011) discuss the opportunities for on-line geoprocessing services to foster multidisciplinary collaboration. The advantages of online geoprocessing tools are many. In particular, shared access to common geospatial data results in considerable savings (Kiehle, 2006), allows users to leverage shared cyberinfrastructure for intensive computing via services such as HUBzero (McLennan and Kennell, 2010), and share workflow elements across different study areas (Yue et al., 2010; Hertel and Villoria, 2012).

Against this background, in order to facilitate the use of these data, we have built a publicly-available, open-source tool that aggregates the data from the grid-cell level to larger geographic aggregates using harvested area and production as alternative weighting schemes. The tool is implemented in GEOSHARE's HUBzero cyberinfrastructure (McLennan and Kennell, 2010) using the statistical language R ({R Core Team}, 2014). GEOSHARE seeks to build a network to improve the availability of data and to facilitate access to spatial data and geoprocessing online tools. Both HUBzero and R are open-source systems, thus saving users costly licenses. Moreover, because the data are entirely handled and processed by GEOSHARE's HUBzero computing resources, users do not require special processing capabilities nor Internet connectivity beyond what is required for ordinary web browsing.

2. Materials and Methods

The AgMIP tool is hosted in GEOSHARE's HUBzero cyberinfrastructure (`https://geoshareproject.org/`). HUBzero (McLennan and Kennell, 2010) is an open source software platform specializing in disseminating simulation and data tools via the world wide web. The HUBzero environment is highly flexible and the only requirement for publishing a tool is that it is written in a language compatible with Linux.

Users access the AgMIP tool @ GEOSHARE using an ordinary Web browser without having to download or compile any code. GEOSHARE is open and users can freely access all the data and simulation tools.

Figure 1 displays the conceptual workflow of the tool separated into three different steps. Each step in figure 1 is implemented as a tab in a graphical user interface shown in figures 2, 3, and 6 . The GGCM output is stored in NetCDF files. Each file is identified by a 12 part file name that specifies crop model, climate model, representative concentration pathway, socio-economic pathway, CO_2 fertilization (yes/no), irrigation (yes/no), a crop, and a time period. For instance:

```
pdssat_hadgem2-es_rcp2p6_ssp2_co2_noirr_yield_mai_annual_2005_2010.nc4
pdssat_hadgem2-es_rcp2p6_ssp2_co2_noirr_yield_mai_annual_2011_2020.nc4
                               .
                               .
                               .
pdssat_hadgem2-es_rcp2p6_ssp2_co2_noirr_yield_mai_annual_2091_2099.nc4
```

are global grids of maize yields (one grid for each year in the period 2005-2099), projected by pDSSAT (Jones et al., 2003; Elliott et al., 2014b), using the climate projections of the the Global Circulation Model HADGEM2-E,under representative concentration pathway RCP2.6, assuming no irrigation and allowing for CO_2 fertilization.

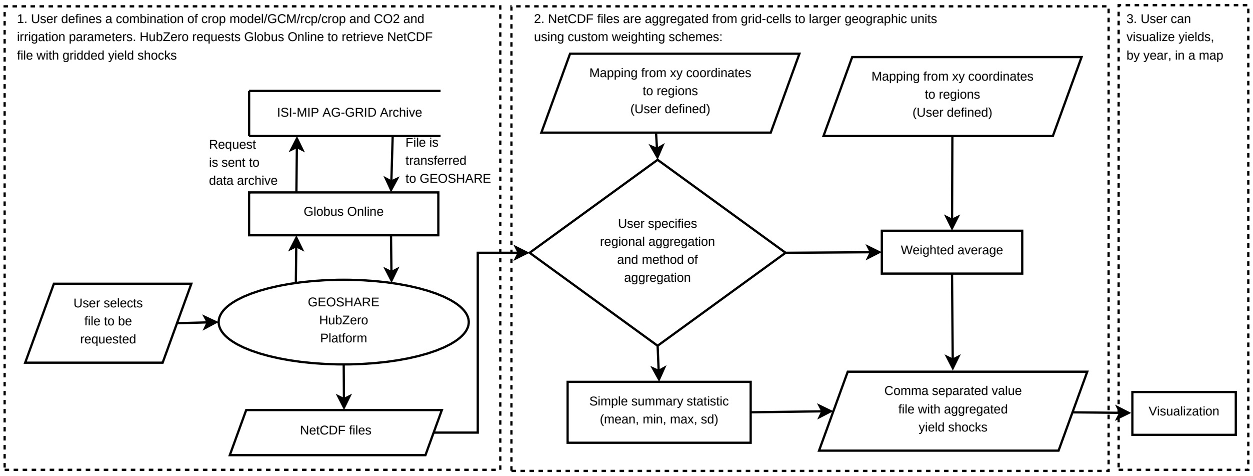
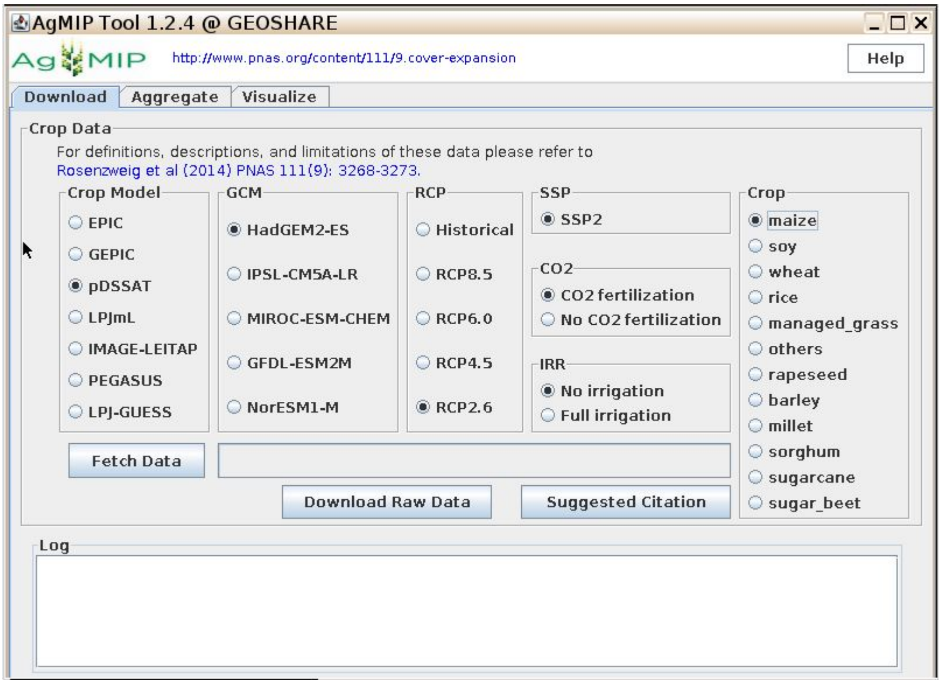


Figure 1: AgMIP tool V1.2. workflow.

In order to retrieve the data, the first step in figure 1 requires the user to select a unique combination of crop model, climate model/scenario, and other choices, which are all presented in the tool's user front-end (figure 2). The user's selections create a character string that matches the file names stored in the the ISI-MIP archive. This character string is used to retrieve all the available years— in most cases, each file stores information on 10 years worth of data— for the selected scenario. GEOSHARE's Hub and the ISI-MIP archive are connected through Globus Online (Foster, 2011), a service that facilitates transfer of large datasets.



For definitions, descriptions, and limitations of these data please refer to Rosenzweig et al (2014) PNAS 111(9): 3268-3273.

Crop Model
- ○ EPIC
- ○ GEPIC
- ● pDSSAT
- ○ LPJmL
- ○ IMAGE-LEITAP
- ○ PEGASUS
- ○ LPJ-GUESS

GCM
- ● HadGEM2-ES
- ○ IPSL-CM5A-LR
- ○ MIROC-ESM-CHEM
- ○ GFDL-ESM2M
- ○ NorESM1-M

RCP
- ○ Historical
- ○ RCP8.5
- ○ RCP6.0
- ○ RCP4.5
- ● RCP2.6

SSP
- ● SSP2

CO2
- ● CO2 fertilization
- ○ No CO2 fertilization

IRR
- ● No irrigation
- ○ Full irrigation

Crop
- ● maize
- ○ soy
- ○ wheat
- ○ rice
- ○ managed_grass
- ○ others
- ○ rapeseed
- ○ barley
- ○ millet
- ○ sorghum
- ○ sugarcane
- ○ sugar_beet

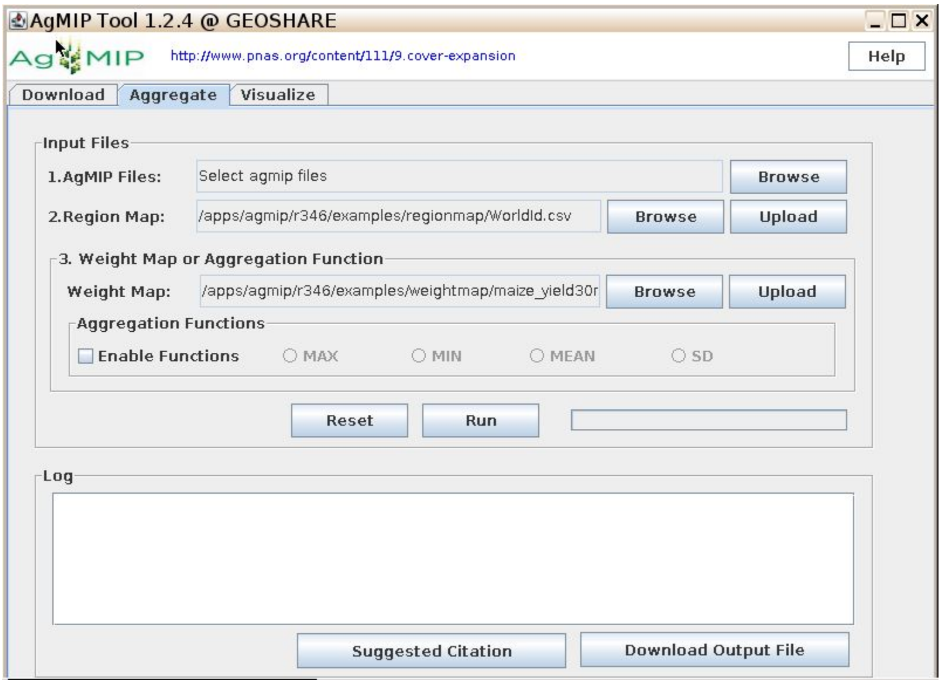
Figure 2: AgMIP tool V1.2. Interface for data selection and retrieval.

Once in GEOSHARE's Hub, the files are stored in a common server workspace. Before each data request, the tool checks whether the data has already been downloaded, and if so, indicates this to the user. This feature avoids downloading the same data more than once. At this point, the user can either download the raw NetCDF files for custom processing on her desktop, or proceed to aggregate the data (step 2 in figure 1 and GUI implementation in figure 3).

2.1. Aggregation

Aggregation is performed by two R functions. The first function uses R NetCDF libraries (Pierce, 2013) to read the data. The second function performs the aggregation. The user has the opportunity to select different aggregation schemes. For example, aggregation from the grid-cell to country level requires a mapping that correlates each latitude and longitude duplet with a unique country name. The mapping schemes are simple comma separated value files. By default, we have included regional mappings for country and country-AEZ regions. However, the user could upload different mapping schemes by employing the download function. Simple guidelines for preparing these data files are in the User's Manual, which can be retrieved from either the description page of the tool, or the "Help" button of the tool's GUI.

In many applications, users need to weight the contribution of each grid-cell to obtain a regional average. The typical measures used as weights are harvested area and production. For the user's convenience, we have converted Monfreda et al. (2008)'s harvested areas from percentages of each 5-minute grid-cell under crop to hectares, assuming a perfect sphere. We have added the physical hectare of the 5 minute grid-



1.AgMIP Files: Select agmip files

2.Region Map: /apps/agmip/r346/examples/regionmap/Worldld.csv

Weight Map: /apps/agmip/r346/examples/weightmap/maize_yield30r

☐ Enable Functions ○ MAX ○ MIN ○ MEAN ○ SD

Figure 3: AgMIP tool V1.2. Aggregation choices.

cells to half degree cells so that they match the spatial resolution of the GGCM files. Further, we have multiplied Monfreda et al. (2008)'s 5-minute grid-cell yields by harvested hectares to obtain production. And as with area, we have added all the production from the 5-min grid-cells to the half degree boxes. As in the case of the regional mappings, the user can also upload custom weighting schemes. Alternatively, the user may decide to use a simple summary statistic (mean, maximum, minimum, and standard deviation). By using if-else statements, the R script decides which function to perform based on these choices. The R functions can be downloaded from the tools description page for further examination and reuse.

2.2. Output and metadata

The output of the R function is a comma-separated-value file with two columns, one identifying the aggregated region(s) and the other the aggregated value (e.g., production weighted average maize yield). We have chosen a CSV format because of its versatility. The tool also produces a detailed description (Suggested Citation in GUI displays) of the operations performed and suggests the citations that should be included when using the tool (see example in the next section).

2.3. Visualization

The last step in the workflow is to visualize the underlying information. At this point, the tool is only able to display aggregations at the country level. We plan on extending the ability of the tool to display other aggregations in future releases.

3. Demonstration & Discussion

The GEOSHARE AgMIP tool is demonstrated using projected maize yields from pDDSAT obtained under HadGEM2 (Collins et al., 2008) climate predictions for representative concentration pathways RCP2.6 and RCP8.5. We include output with and without CO_2 fertilization. pDDSAT (Jones et al., 2003; Elliott et al., 2014a) is a modification of the DSSAT crop model that runs in global grids, given the basic information on biophysical attributes. The user of these yield shocks is advised to consult Rosenzweig et al. (2014) for a discussion of the different models and attributes.

Figure 2 shows the front-end of AgMIP tool V.1.2 with the relevant options checked. Figure 3 displays the "Aggregation" tab. Beyond the included default mapping and weighting schemes, the tool is intended to give the user flexibility regarding choices of aggregation. To demonstrate such flexibility we created two additional regional mapping schemes. One maps each coordinate pair to a unique aggregate unit, which we label "World." The other maps each coordinate pair into one of two categories: low latitude (regions within the tropics) and mid- latitude (between 23.5^o and 66^o).

Figure 4 displays the global production-weighted average maize yield from 2005 to 2099. To gain some perspective on the size of this job, consider that each RCP-CO2 fertilization scenario of pDDSAT is stored in 9 NetCDF array files, each containing 10 years worth of data. And for each year there are 720 X 360 = 259,200 grid-cells. Thus, for the four sets of outputs illustrated in figure 3 (2 RCPs, with and without CO_2), the tool downloads 36 NetCDF files and processes 324 (36×9) grids to aggregate to a single summary annual statistic. In the process, the final user is shielded from all the technical barriers implied by learning new download platforms and specialized formats, as well as pre-processing of the weighting schemes. Moreover, this can be all performed trough any standard web browser without any special bandwidth requirement.

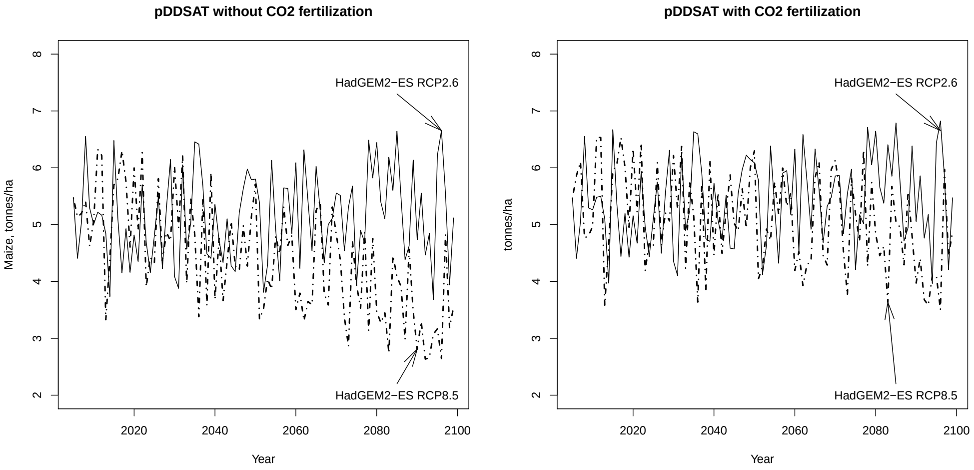


Figure 4: Future maize yields (world average, weighted by grid-cell level production) projected by pDDSAT using HadGEM2-ES future climates under representative concentration pathways scenarios RCP2.6 and RCP8.5.

In figure 5 we demonstrate an alternative mapping scheme in which the world is divided into low and middle latitudes as explained above. In addition, we demonstrate the effects of using production weights, area weights, and simple averages.

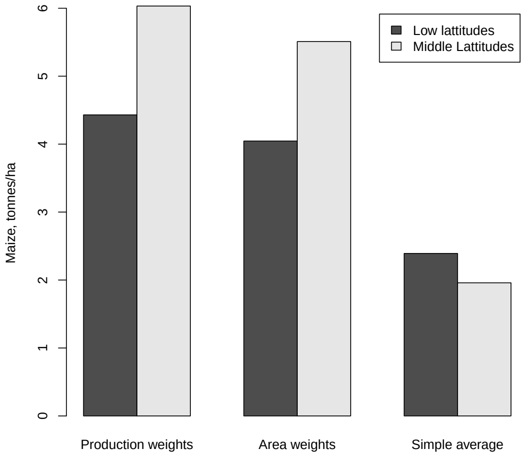
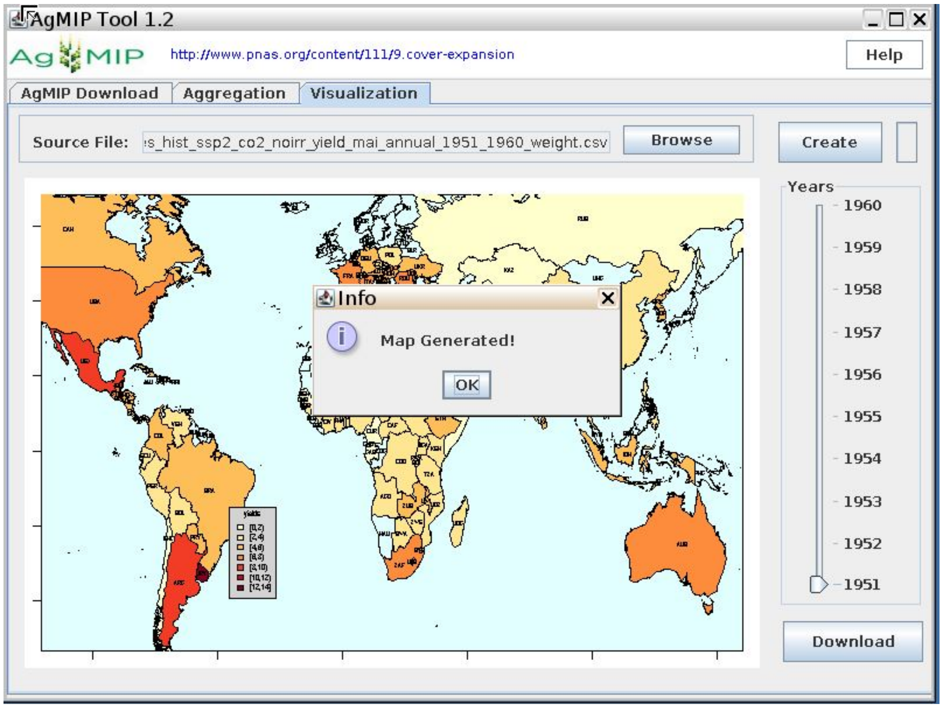


Figure 5: Average maize yields in low and middle latitudes in the year 2050 under alternative weighting schemes. Output from pDDSAT under HadGEM2-ES RCP2.6.

Figure 6 shows the visualization tab. As mentioned above, at this point this tool is capable of displaying only country level information. The slide on the right of the tool allows for the selection of the different years in the aggregation.

4. Conclusions

Access to spatial datasets by non specialists is hindered by technical difficulties involving software and data formats as well as the need for strong Internet bandwidth and storage capacity. Here we discuss a GEOSHARE tool that expands access to the outputs from the AgMIP Global Gridded Crop Model Intercomparison Project to the broader scientific community who can benefit from these data, but who may lack the resources to gain access to them.



Source File: :s_hist_ssp2_co2_noirr_yield_mai_annual_1951_1960_weight.csv

Map Generated!

Figure 6: AgMIP tool V1.2. Visualization.

References

Collins, W., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Hinton, T., Jones, C., Liddicoat, S., Martin, G., OConnor, F., Rae, J., others, 2008. Evaluation of the HadGEM2 model. Hadley Cent. Tech. Note 74.

Craglia, M., de Bie, K., Jackson, D., Pesaresi, M., Remetey-Flpp, G., Wang, C., Annoni, A., Bian, L., Campbell, F., Ehlers, M., van Genderen, J., Goodchild, M., Guo, H., Lewis, A., Simpson, R., Skidmore, A., Woodgate, P., Dec. 2011. Digital earth 2020: towards the vision for the next decade. International Journal of Digital Earth 5 (1), 4–21.
URL `http://dx.doi.org/10.1080/17538947.2011.638500`

Elliott, J., Deryng, D., Müller, C., Frieler, K., Konzmann, M., Gerten, D., Glotter, M., Flrke, M., Wada, Y., Best, N., Eisner, S., Fekete, B. M., Folberth, C., Foster, I., Gosling, S. N., Haddeland, I., Khabarov, N., Ludwig, F., Masaki, Y., Olin, S., Rosenzweig, C., Ruane, A. C., Satoh, Y., Schmid, E., Stacke, T., Tang, Q., Wisser, D., Mar. 2014a. Constraints and potentials of future irrigation water availability on agricultural production under climate change. Proceedings of the National Academy of Sciences 111 (9), 3239–3244.

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., Foster, I., 2014b. The parallel system for integrating impact models and sectors (pSIMS). Environmental Modelling & Software, 1–8.
URL `http://www.sciencedirect.com/science/article/pii/S1364815214001121`

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Bchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T., Izaurralde, R. C., Müller, N. D., Ray, D. K., Rosenzweig, C., Ruane, A. C., Sheffield, J., Jul. 2014c. The global gridded crop model intercomparison: data and modeling protocols for phase 1 (v1.0). Geosci. Model Dev. Discuss. 7 (4), 4383–4427.
URL `http://www.geosci-model-dev-discuss.net/7/4383/2014/`

Foster, I., 2011. Globus online: Accelerating and democratizing science through cloud-based services. Ieee Internet Computing 15 (3), 70–73.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber`

Hertel, T. W., Britz, W., Diffenbaugh, N. S., Ramankutty, N., Villoria, N., Oct. 2010. A global, spatially explicit, open source data base for analysis of agriculture, forestry, and the environment: Proposal and institutional considerations. Report to the UK science advisor, Department of Agricultural Economics, Purdue University.

Hertel, T. W., Villoria, N., Apr. 2012. GEOSHARE: Geospatial open source hosting of agriculture, resource and environmental data for discovery and decision making. GPRI Digital Library.
URL `http://docs.lib.purdue.edu/gpridocs/7`

Hofer, B., Sep. 2014. Uses of online geoprocessing technology in analyses and case studies a systematic analysis of literature. International Journal of Digital Earth 0 (ja), 1–21.
URL `http://dx.doi.org/10.1080/17538947.2014.962632`

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., Ritchie, J. T., Jan. 2003. The DSSAT cropping system model. European Journal of Agronomy 18 (34), 235–265.
URL `http://www.sciencedirect.com/science/article/pii/S1161030102001077`

Kiehle, C., Dec. 2006. Business logic for geoprocessing of distributed geodata. Computers & Geosciences 32 (10), 1746–1757.
URL `http://www.sciencedirect.com/science/article/pii/S0098300406000756`

McLennan, M., Kennell, R., Mar. 2010. HUBzero: A platform for dissemination and collaboration in computational science and engineering. Computing in Science & Engineering 12 (2), 48–53.
URL `http://scitation.aip.org/content/aip/journal/cise/12/2/10.1109/MCSE.2010.41`

Monfreda, C., Ramankutty, N., Foley, J. A., Mar. 2008. Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Global Biogeochemical Cycles, 1:19.

Monfreda, C., Ramankutty, N., Hertel, T., 2009. Global agricultural land use data for climate change analysis. In: Hertel, T. W., Rose, S. K., Tol, R. S. (Eds.), Economic Analysis of Land Use in Global Climate Change Policy. Routdlege, London and New York, pp. 33–49.

Nelson, G. C., Valin, H., Sands, R. D., Havlk, P., Ahammad, H., Deryng, D., Elliott, J., Fujimori, S., Hasegawa, T., Heyhoe, E., Kyle, P., Lampe, M. V., Lotze-Campen, H., dCroz, D. M., Meijl, H. v., Mensbrugghe, D. v. d., Müller, C., Popp, A., Robertson, R., Robinson, S., Schmid, E., Schmitz, C., Tabeau, A., Willenbockel, D., Mar. 2014. Climate change effects on agriculture: Economic responses to biophysical shocks. Proceedings of the National Academy of Sciences 111 (9), 3274–3279.
URL `http://www.pnas.org/content/111/9/3274`

Pierce, D., 2013. ncdf4: Interface to unidata netCDF (version 4 or earlier) format data files.
URL `http://dwpierce.com/software`

Piontek, F., Müller, C., Pugh, T. A. M., Clark, D. B., Deryng, D., Elliott, J., Gonzlez, F. d. J. C., Flrke, M., Folberth, C., Franssen, W., Frieler, K., Friend, A. D., Gosling, S. N., Hemming, D., Khabarov, N., Kim, H., Lomas, M. R., Masaki, Y., Mengel, M., Morse, A., Neumann, K., Nishina, K., Ostberg, S., Pavlick, R., Ruane, A. C., Schewe, J., Schmid, E., Stacke, T., Tang, Q., Tessler, Z. D., Tompkins, A. M., Warszawski, L., Wisser, D., Schellnhuber, H. J., Mar. 2014. Multisectoral climate impact hotspots in a warming world. Proceedings of the National Academy of Sciences 111 (9), 3233–3238.
URL `http://www.pnas.org/content/111/9/3233`

{R Core Team}, 2014. R: A language and environment for statistical computing.
URL `http://www.R-project.org`

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A. M., Schmid, E., Stehfest, E., Yang, H., Jones, J. W., Mar. 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.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., Antle, J. M., Nelson, G. C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J. M., Mar. 2013. The agricultural model intercomparison and improvement project (AgMIP): Protocols and pilot studies. Agricultural and Forest Meteorology 170, 166–182.
URL `http://www.sciencedirect.com/science/article/pii/S0168192312002857`

Stasch, C., Scheider, S., Pebesma, E., Kuhn, W., Jan. 2014. Meaningful spatial prediction and aggregation. Environmental Modelling & Software 51, 149–165.
URL `http://www.sciencedirect.com/science/article/pii/S1364815213001977`

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., Schewe, J., Mar. 2014. The inter-sectoral impact model intercomparison project (ISIMIP): Project framework. Proceedings of the National Academy of Sciences 111 (9), 3228–3232.
URL `http://www.pnas.org/content/111/9/3228`

Yue, P., Gong, J., Di, L., Yuan, J., Sun, L., Sun, Z., Wang, Q., Dec. 2010. GeoPW: Laying blocks for the geospatial processing web. Transactions in GIS 14 (6), 755–772.
URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2010.01232.x/abstract`