Improving the Computational Efficiency of Downscaling GCM data for Use in SWAT

REU Site: Interdisciplinary Program in High Performance Computing

Christopher Evans[1], Abigail Gartrell[2], Lauren Gomez[3], Moise Mouyebe[4], Darius Oxley[5],
Graduate assistant: Sai Kumar Popuri[5], Faculty mentor: Nagaraj K. Neerchal[5],
Client: Amita Mehta[6]

[1]Department of Mathematics, Hampden-Sydney College
[2]Department of Mathematics, University of Maryland, College Park
[3]Department of Mathematics, California State Polytechnic University, Pomona
[4]Department of Mathematics, University of Michigan, Flint
[5]Department of Mathematics and Statistics, University of Maryland, Baltimore County,
[6]Joint Center for Earth Systems Technology (JCET) and Geography and Environmental Systems,
University of Maryland, Baltimore County

Technical Report HPCF–2014–12, www.umbc.edu/hpcf > Publications

Abstract

This project creates software tools to streamline the computational procedures to generate high-resolution weather parameters from low resolution Global Climate Models (GCM) to be input into the Soil and Water Assessment Tool (SWAT) and to visualize GCM temperature and precipitation data as well as SWAT outputs of crop data. Data from GCMs have relatively low spatial resolution ($\sim$ 100km x 100km), which needs to be downscaled to a higher resolution to match the resolution of the observed data so that both the data sets can be input into SWAT. We consider several decades of historical simulations from GCMs, and surface-based observations of historical temperature and precipitation data over the Missouri River Basin (MRB). The downscaling method involves two steps: i) bilinear interpolation to fill in values at higher resolution and ii) Linear or Tobit regression between the GCM and the observed data to accurately capture the features of observed data in the GCM temperature and precipitation at high resolution. This downscaled data is then used to generate forecasts, which are used as inputs for SWAT. Based on these statistical downscaling methods, the SWAT outputs are compared. Since the data is large, we focus on maximizing computational efficiency through parallelization of the data generation, model fitting and forecasting. A graphical R interface is developed to facilitate the modeling and visualization components at each step. The interface allows climate models and SWAT outputs to be more easily compared for different scenarios, thus help assess climate variability and its impact on crop yields over MRB.

Key Words: SWAT, GUI, Tobit Regression, Downscaling, GCM, MRB

1 Introduction

The Missouri River Basin is the largest river basin in the United States, covering more than 500,000 square miles (approximately 1,280,000 sq. km) and includes parts of ten U.S. states and two Canadian provinces. The river basin provides drinking water, irrigation, hydro-electricity and a habitat for a range of fish and wildlife. It is also a very important food-producing region, producing approximately 46% of U.S. wheat, 22% of its grain corn, and 34% of its cattle. Approximately 90% of the basin's cropland is not irrigated and is entirely dependent on precipitation. Such a large percentage of agricultural and hydrological output makes a focus on the basin a necessity for assessing climate impact in terms of both water and crop yields in order to develop management strategies for the river basins and watersheds of the U.S.

Climate is the average weather conditions prevailing in an area over a long period of time. Such weather conditions include atmospheric pressure, temperature, humidity, wind, and precipitation and are generally averaged over a period of 30 years [3]. Climate variability and change impact fresh water availability and agriculture. With long-term, quality observations available for validation, the Missouri River Basin is an excellent region to understand the effect of climate variability, particularly temperature and precipitation, on water and crop yields. Therefore, Climate model simulations, downscaled to 10km x 10km grid resolution from the original 100km x 100km, are used in conjunction with a hydrology and crop model, the Soil Water Assessment Tool (SWAT), to simulate water and crop yields over the Missouri River Basin. The downscaling and the SWAT simulations are computationally intense and require a large amount of data management. The primary objective of this project is to streamline the computational procedure leading up to SWAT simulations.

In this project we focus on: (1) generating high-resolution weather parameters (temperature and precipitation) by downscaling Global Climate Models (GCM) data and inputting these parameters into SWAT, (2) parallelizing model fitting and forecasting and (3) developing a Graphical User Interface (GUI) tool to perform calculations on any given GCM data set, and visualizing temperature and precipitation data. All the components of this project are developed and tested with a single climate model simulation; however, this procedure can be used to conduct experiments for a variety of climatic conditions simulated by a number of climate models with improved computational efficiency and graphical aid.

The rest of this paper is organized as follows. Section 2 gives a more detailed look into the background information of the project. Section 3 describes the statistical and computing methods used. Lastly, Section 4 gives some concluding remarks.

2 Background

Long-term independent records from weather stations, satellites, ocean buoys, and many other data sources confirm that natural climate variability and anthropogenic climate change in regional temperature, precipitation, winds and humidity have significant impacts on fresh water availability and agricultural output. For instance, according to the 2014 National Climate Assessment (NCA) report [2], people in the Midwest region will have to get used to extreme heat, heavy rainfalls and flooding ultimately affecting infrastructure, transportation, forestry, air and water quality and agriculture as well. Also, those living in the Great Plains region will experience rising temperature resulting in an increasing demand for fresh water and energy. In order to better assess these vital challenges posed by the changing climate conditions, a multi-institute (CRCES, Texas A&M, UMBC-JCET, NDMC) project supported by the US Department of Agriculture-National Institute for Food and Agriculture (USDA-NIFA) was created to assess the impacts of natural decadal climate variability on water availability and crop products in the Missouri River Basin (MRB) with future goal for expansion to other geographic regions and for various climatic conditions.

The client-team from UMBC-JCET has used both daily and monthly low resolution data for precipitation (pr), maximum and minimum temperature (tasmax and tasmin respectively) provided by the two Global Climate Models HadCM3 [8] and MIROC5 [12] to generate high resolution downscaled data needed to run SWAT [5]. SWAT is a modeling tool that is used to simulate agricultural yields and hydrological impact of climate variability and change. The team has conducted a comparison analysis of the downscaled coefficients interpolated from both daily and monthly GCM data, and has concluded that even though the coefficients calculated from the monthly GCM data are computationally easy to obtain as far as the run-time is concerned, they are less accurate compared to those obtained from daily GCM data when it comes to making climate forecasts. As a result, the team has decided to only use the daily GCM data for downscaling purposes.

For the past year or so the UMBC-JCET team has focused on the MIROC5 data, and it has come to realize that even though the downscaled temperature forecasts are close to the observed temperature, the same is not true for precipitation. This observation has forced the team to look for ways to improve the quality of downscaling, and presently the following three techniques are being explored: First, using Surface Level Pressure (SLP) as a covariate while performing regression. Second, using a two-component *Logit* [4] model instead of a linear regression model and finally, using a *Tobit* [1] model. We have used the Tobit model in our project.

3 Methods and Results

3.1 Downscaling

For our study, we consider daily observed data for 57 years (1949-2005). Maurer [7] provides these data for MRB, indexed by latitude and longitude. The data from the GCM MIROC5 is available for 1859-2005, albeit at a lower resolution. Both MIROC5 and Maurer data sets include daily precipitation and temperature (daily minimum and maximum), which are needed as input for running SWAT and to obtain crop and water yields for subregions of MRB. MIROC5 follows a 365-day calendar and Maurer is on the regular calendar. In order to bring the MIROC5 data on to the regular calendar, data for the last day in February for each leap year is repeated.

The model data is at a low resolution, as it looks at a larger general area than the observed data. For instance, MIROC5 looks at the Missouri River Basin in squares of 100km x 100km while the observed data covers the basin in a grid of 10km x 10km. In order to generate enough data points to compare to the daily observed data, we must downscale. The downscaling methods include statisitcal methods of interpolation and regression. Specifically, we use bilinear interpolation, Equation 3.1, in order to compute enough data points from the model data to compare to the observed data. Bilinear interpolation, modeled in Figure 3.1, allows us to take a region of model data, say a square of certain latitudes and longitudes. Then, knowing the climate data at the four known points we are able to approximate the value within certain locations on the line segements of the square. We do this twice, between the two different sets of points. Then we interpolate again in order to find specfic points in between the new interpolated points. We are able to calculate points within the square of the model data.

$$\begin{aligned} f(x,y) = \frac{1}{(x_2-x_1)(y_2-y_1)} &(f(x_1,y_1)(x_2-x)(y_2-y) + f(x_2,y_2)(x-x_1)(y_2-y) \\ &+ f(x_1,y_2)(x_2-x)(y-y_1) + f(x_2,y_2)(x-x_1)(y-y_1)) \end{aligned} \tag{3.1}$$

Interpolated model data and observed data provide us the data structure needed to apply regression methodology. The regression methodology is defined as follows:

$$y_{sij} = \beta_{0sij} + \beta_{1sij}x_{sij} + \epsilon_{sij} \tag{3.2}$$

Here, $\{y_{sij}\}$ is the 57×1 vector of observed data for the period 1949-2005, $\{x_{sij}\}$ is the corresponding vector of model data, and $\{\epsilon_{sij}\}$ is the vector of errors at location $\{s\}$ for the day i of month j. Thus, we obtain slope and intercept estimates corresponding to each combination of location, day and month. However, because there are many days where the precipitation is zero, the assumption of linearity is untenable. Therefore, the downscaled model data for precipitation shows a discrepancy

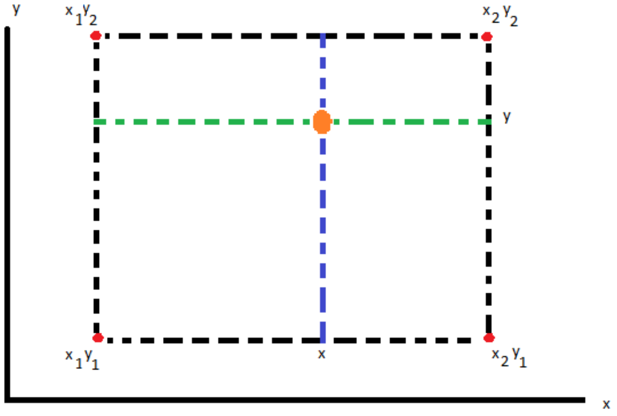


Figure 3.1: Schematic of bilinear interpolation

with the observed data when simple linear regression is used. Since the presence of zeros in the precipitation data prevents the use of linear regression, a Tobit regression is considered, in order to generate regression coefficients. For more information, see Popuri [9].

Next step is to evaluate the performance of the above estimated regression equation in producing valid downscaled data. The approach normally taken by the geoscientists involves using hindcast data (retrospective forecast - see Taylor [11] for more information) as predictor and compare the predictions to correpsonding observed values. We use the hindcast data from 1981-1990 as predictor to generate predictions for the same period. Figure 3.2 shows the workflow of the model fitting and prediction process. The hindcast data, like the simulated model data, is at low resolution. Using bilinear interpolation we generate high resolution hindcast data. Using the daily regression coefficients and the hindcast data as predictor, we calculate forecasted data, which is then used as input to SWAT. Outputs from SWAT runs using the forecasted data generated by the downscaling method are then compared to the outputs generated using the observed data for the years 1981-1990 in order to evaluate the performance of the downscaling procedure.

3.2 Parallelization

We use the statistical programming language R [10] as it is a widely used open source language with a large number of sophisticated add-on packages. One of our goals includes making the downscaling routine more efficient. To this end, we decided to parallelize the generation of the ensemble level data and the regression routine. We have implemented a parallelized version of the code in the maya cluster [6] using the R package ”snow” with ”socket” communication.

```
start.time <- Sys.time()
np <- 32
cluster <- makeSOCKcluster(rep("localhost", np))
E <- parLapplyLB(cluster, dta, slr_fit)
stopCluster(cluster)
end.time <- Sys.time()
```

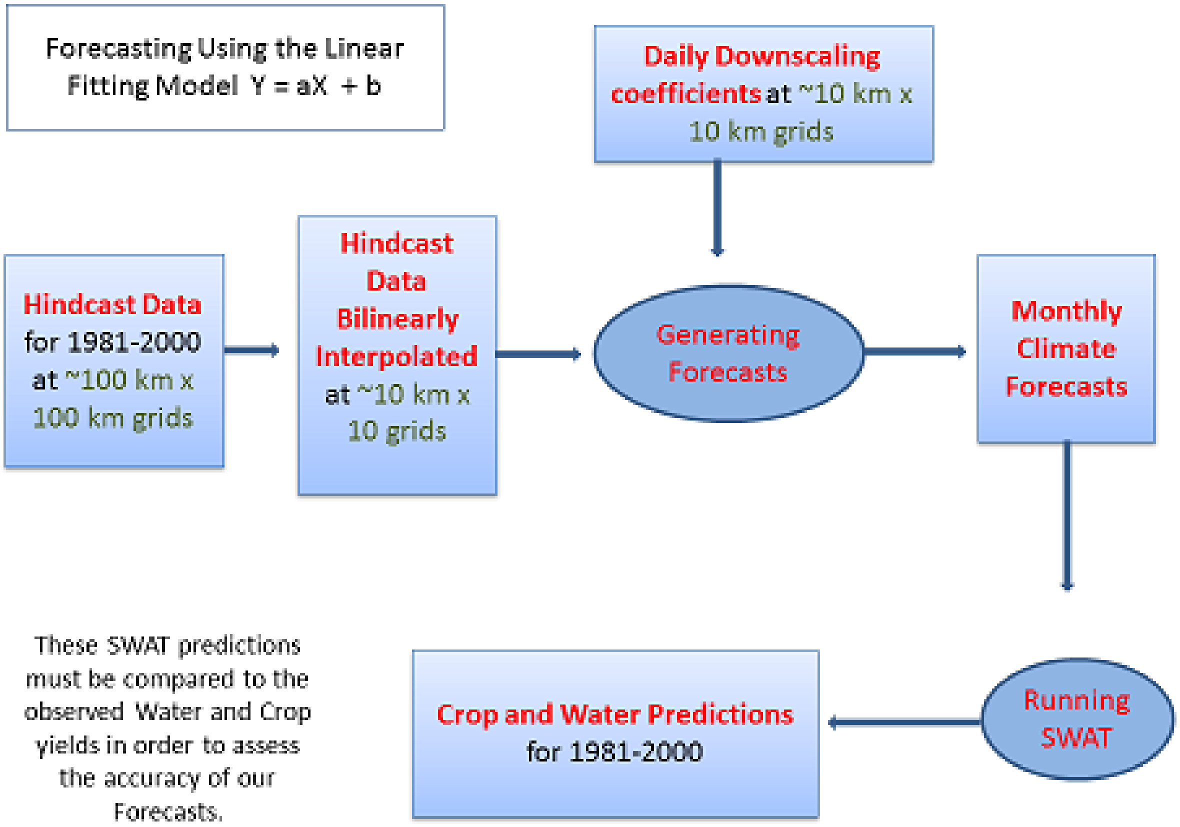

Forecasting Using the Linear Fitting Model Y = aX + b
Daily Downscaling coefficients at ~10 km x 10 km grids
Hindcast Data for 1981-2000 at ~100 km x 100 km grids
Hindcast Data Bilinearly Interpolated at ~10 km x 10 grids
Generating Forecasts
Monthly Climate Forecasts
These SWAT predictions must be compared to the observed Water and Crop yields in order to assess the accuracy of our Forecasts.
Crop and Water Predictions for 1981-2000
Running SWAT

Figure 3.2: Flow chart depicting generation of forecast data for SWAT

Table 3.1: Timing of Regression in Serial and Parallel (seconds)

1 node(serial)	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1372.271	805.687	394.064	240.740	129.991	137.532

In the code above, the number of nodes is 32 while the function 'parLapplyLB' passes the data 'dta', which is a list of observed and model data, along with specific longitudes and latitudes, to the maya cluster and performs the regression function 'slr_fit'. The data is then returned in the form of lists with the daily regression coefficients. We use a similar routine for the generation of forecasted ensemble level data. In this routine, each compute node is configured to work on a different ensemble. We have timed the parallelized regression routine and the generation of forecasted data for one month and noticed significant decrease in the run time.

As can be seen from Table 3.1, regression run time declines until 32 nodes. Figure 3.3 graphically shows the speedup behavior as the number of nodes increase. When we reach 32 nodes however, the speedup begins to decline. We suspect that the reason behind this is inter-nodal communication overhead. Also, we parallelized using the R package 'snow' which gives us less control over the nodes and processes. The issue of the decrease in speedup needs further investigation.

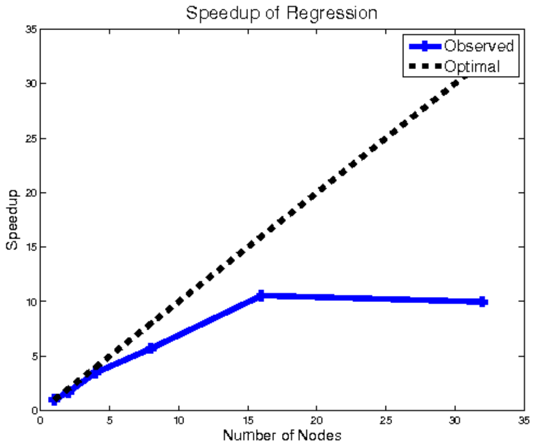


Figure 3.3: Speed up plot of the parallel regresssion routine. The black dashed line represents the optimal solution while the blue line represents what was actually observed.

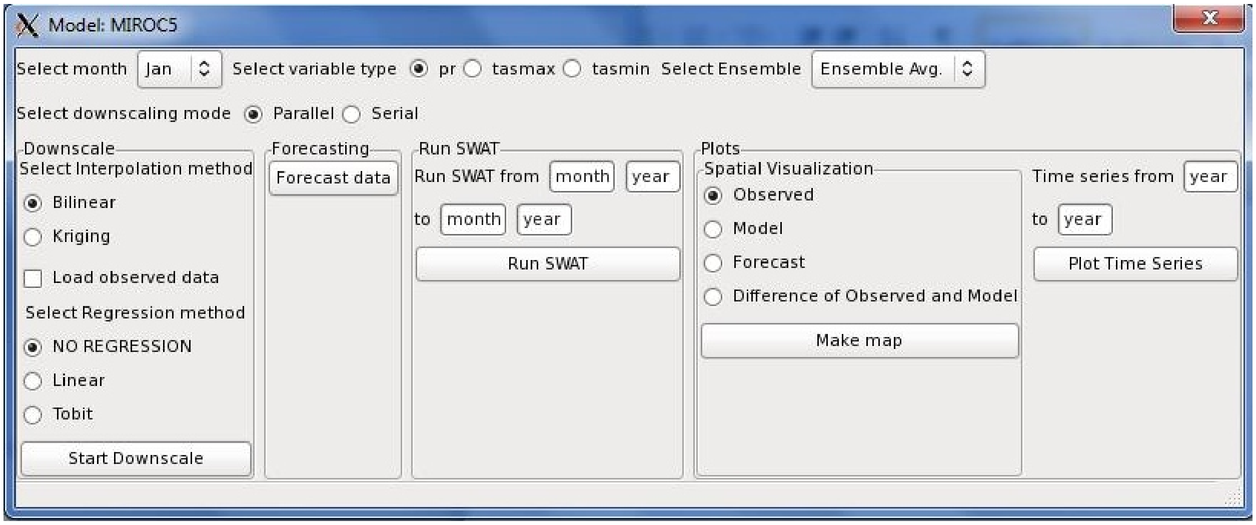
3.3 Graphical User Interface

A Graphical User Interface(GUI) tool, Figure 3.4, was developed to perform various statistical calculations including downscaling and regression on any given Global Climate Model data set. This interactive device allows the user to streamline the process of downscaling GCM data, performing regression, forecasting, inputting the data into SWAT and visualizing the output. The different models and ensembles can more quickly be compared to one another. The modeling process was generalized so that it can be implemented for any GCM, whether it is from MIROC5 or HadCM3. The visualization methods include maps that display average monthly data, the differences between monthly averages, and the average over several months for all model and observed data. Additionally, time series plots are available for the user to investigate monthly and yearly averages of each data set and how they compare with the others. The yearly averages time series allows the user to select a range of dates and then plot the average over all the locations in MRB for all the months in a specific year. Thus, it is easier to compare different model data to the observed data and provides analysis of the most effective methods.

Before performing any statistical operation, the user can visualize the raw data of both the observed and model data in either a spatial or temporal manner. For each month and each variable type, the user can use Bilinear Interpolation to create high-resolution data. Furthermore, they have the options of Linear Regression, Tobit Regression (for precipitation only), or not to use regression.

After downscaling, the user can view the monthly averages of precipitation, maximum or minimum temperature on a map of the Missouri River Basin. Figure 3.5 is an example of one such map, which allows the user to view rainfall in the MRB region. This provides visualisation of the data, so the user can become more familiar with it. Next, the user can use the downscaled coefficients created in the previous step to forecast the precipitation and temperature, as described Section 3.1. Once the data is present, the user has the option to view a time series of the different data, like in Figure 3.6. This enables expedited analysis of the downscaling methods before the data is sent to SWAT.

Finally, the forecasted data is used as an input to SWAT. The user can easily see the expected water and crop yields based on selected parameters such as the chosen GCM, ensemble number, interpolation type, and regression method. Thus, the GUI tool streamlines the downscaling process and the regression routines are parallelized in order to make the computations more efficient.



Model: MIROC5

Select month Jan

Select variable type ● pr ○ tasmax ○ tasmin

Select Ensemble Ensemble Avg.

Select downscaling mode ● Parallel ○ Serial

Downscale

Select Interpolation method

● Bilinear

○ Kriging

□ Load observed data

Select Regression method

● NO REGRESSION

○ Linear

○ Tobit

Forecasting

Run SWAT

Run SWAT from month year

to month year

Plots

Spatial Visualization

● Observed

○ Model

○ Forecast

○ Difference of Observed and Model

Time series from year

to year

Figure 3.4: GUI Screenshot

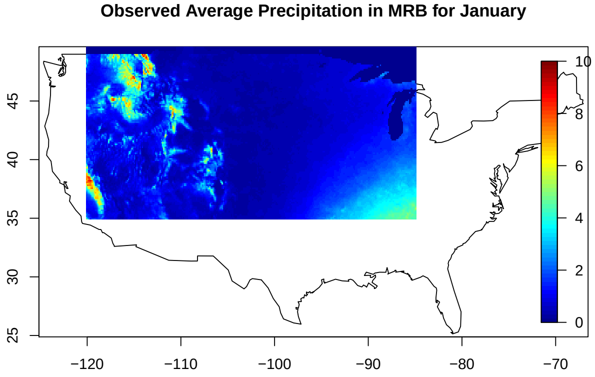


Figure 3.5: The average observed rain in January for each location over 1949 to 2005

4 Concluding Remarks

In this paper we discussed generalizing the downscaling process to handle any Global Climate Model. The low resolution data was successfully downscaled through bilinear interpolation and both linear and Tobit regression were parallelized using the maya cluster and the R package 'snow' to significantly improve the efficiency of downscaling. We have also produced a GUI to make a user friendly interface for streamlining the process of downscaling of any GCM data. The GUI is also capable of generating forecasts, run SWAT, and visualize observed model or forecasted data.

Our research has given us insight into ideas for future research on this topic. Since streamlining allows the comparison of more GCMs, further research would help in discovering previously unob-

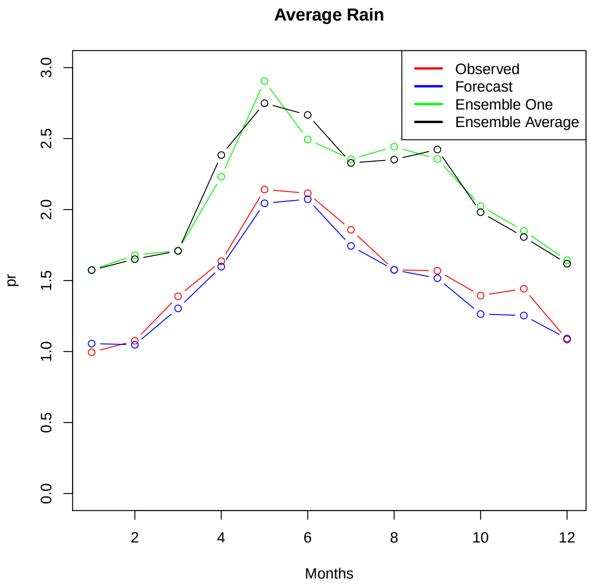


Figure 3.6: The average rain in each month for the years 1981 to 2000 over the Missouri River Basin for each model.

served relations in data. Parallelization allows more data to be processed, opening up larger areas of land to be analyzed under the high resolution model. Another potential topic to explore is the method of interpolation called Kriging. Kriging takes into account the topography of the land and uses the covariances among the different locations and might help predict better data at different locations in the Missouri River Basin. Producing the input for and running SWAT was a challenge as SWAT is a Fortran based tool; however, we were close to formatting our downscaled data into the proper SWAT .pcp and .tmp files so that we could run SWAT with the high resolution data. We could also run SWAT with observed data and model data in order to compare agricultural yields. Being able to run SWAT with our more precise data would allow us to give better predictions of crop and water output and help farmers with that information.

Acknowledgments

These results were obtained as part of the REU Site: Interdisciplinary Program in High Performance Computing (www.umbc.edu/hpcreu) in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County (UMBC) in Summer 2014. This program is funded jointly by the National Science Foundation and the National Security Agency (NSF grant no. DMS–1156976), with additional support from UMBC, the Department of Mathematics and Statistics, the Center for Interdisciplinary Research and Consulting (CIRC), and the UMBC High Performance Computing Facility (HPCF). HPCF is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258 and CNS–1228778) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from UMBC. Authors would like to acknowledge Joint Center for Earth Systems Technology(JCET) and the Center for Research in the Changing Earth Systems (CRCES) for their support. In partcular, we would like to thank Katherin Mendoza for help with the SWAT runs. Co-author Darius Oxley was supported, in part, by the UMBC National Security Agency (NSA) Scholars Program through a contract with the NSA.

Graduate assistant Sai Kumar Popuri was supported during Summer 2014 by UMBC.

References

[1] T. Amemiya. *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusetts, 1985.

[2] U. S. National Climate Assessment. Climate change impacts in the united states, 2014. This report can be found at the URL http://nca2014.globalchange.gov.

[3] A.P.M. Baede. Intergovernmental panel on climate change. http://www.grida.no/publications/other/ipcc_tar/?src=/climate/ipcc_tar/wg1/518.htm.

[4] D. Fletcher, D. Mackenzie, and E. Villouta. Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics*, 12:45–54, 2005.

[5] P.W. Gassman, M.R. Reyes, C.H. Green, and J.G. Arnold. The soil and water assessment tool: Historical development, applications, and future research directions. http://www.card.iastate.edu/environment/items/asabe_swat.pdf.

[6] UMBC HPCF. UMBC high performance computing facility, 2014. http://www.umbc.edu/hpcf/systems.html.

[7] E.P. Maurer, A.W. Wood, J.C. Adam, and D.P. Lettenmaier. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states. *Journal of Climate*, 15(22):3237–3251, 2002.

[8] V.D. Pope, M.L. Gallani, P.R. Rowntree, and R.A. Stratton. The impact of new physical parametrizations in the Hadley centre climate model — HadAM3. *Climate Dynamics*, 2000.

[9] S.K. Popuri, A. Mehta, and N.K. Neerchal. Forecasting daily precipitation over Missouri River Basin using simulated data by MIROC5.

[10] R Core Team. R: A Language and Environment for Statistical Computing, 2014. http://www.R-project.org.

[11] K.E. Taylor, R.J. Stouffer, and G.A. Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[12] M. Watanabe, T. Suzuki, et al. Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *Journal of Climate*, 2010.