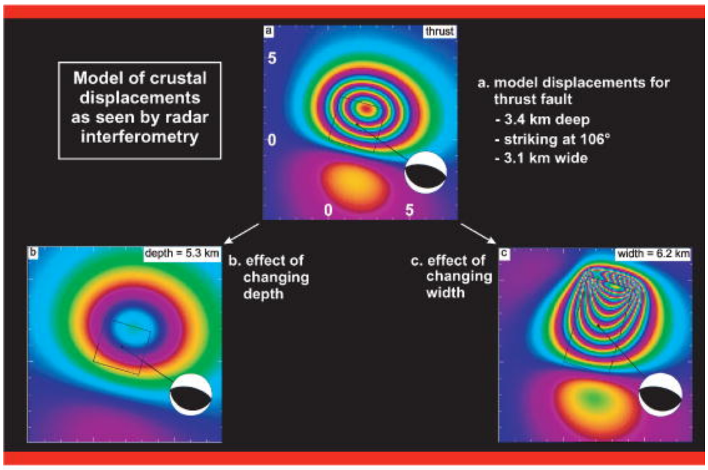
Volume 33 Issue 4 May 2007 ISSN 0098-3004

ELSEVIER

COMPUTERS & GEOSCIENCES

An International Journal

Editor-in-Chief
Eric C. Grunsky
Geological Survey of Canada


Model of crustal displacements as seen by radar interferometry
a
thrust
5
0
0
5
a. model displacements for thrust fault
- 3.4 km deep
- striking at 106°
- 3.1 km wide
b
depth = 5.3 km
b. effect of changing depth
c. effect of changing width
c
width = 6.2 km

See Editor's Home Page
http://www.iamg.org/CGEditor/index.htm
to download codes for programs discussed

ELSEVIER

Computers & Geosciences 33 (2007) 457–464

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Automating regional descriptive statistic computations for environmental modeling

Satoshi Hirabayashi, Charles Kroll*

State University of New York College of Environmental Science and Forestry, 312 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210, USA

Received 19 February 2006; received in revised form 31 May 2006; accepted 22 June 2006

Abstract

A GIS toolset was developed to support the extraction of a variety of input variables for environmental models. The developed toolset allows the automated processing of large amounts of raster data over polygon data. A case study was performed in a region centered on eastern Tennessee and western North Carolina. Using the developed GIS toolset, topographic, geologic, and climatic characteristics for watersheds corresponding to 35 United States Geological Survey streamflow gauging sites were derived by processing approximately 1500 raster datasets. The developed GIS toolset greatly reduced the time and effort needed to process the GIS data, and provides a useful tool for a wide variety of environmental applications. The developed toolset is freely available for download, and a tutorial has been created.

Keywords: GIS; Batch process; Raster; Environmental modeling

1. Introduction

Environmental models simulate environmental conditions and processes, both of which have spatial attributes. Geographic Information Systems (GIS) can be used to define and describe spatial properties of the environment. Although they evolved separately, GIS and environmental modeling have recently enhanced their synergies to each other, and GIS can now serve as a fundamental data and analysis framework for environmental models (Maidment, 1996; Rao et al., 2000). GIS is commonly utilized to prepare data, to extract model parameters, and to visualize model results in a wide variety of environmental studies.

In environmental modeling, regions of interest such as state or county boundaries, watershed boundaries, or other land parcels are often represented as polygons in vector-formatted GIS data. On the other hand, values of interest such as elevation, land cover, and climatic data are often represented by grid cells in raster-formatted GIS data. By overlaying these two types of GIS data, descriptive statistics can be computed from values assigned to grids that fall within each polygon. Examples of such statistics in environmental studies are predominant aspect class (majority), maximum precipitation (maximum), average elevation (mean), and number of unique land covers (variety). A large amount of input data derived from raster datasets is often required in a variety of environmental models

*Tel.: +1 315 470 6699; fax: +1 315 470 6958.
E-mail addresses: shirabay@syr.edu (S. Hirabayashi), cnkroll@esf.edu (C. Kroll).

0098-3004/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.cageo.2006.06.013

(Pullar and Springer, 2000; He, 2003; Bedient et al., 2000; Neary et al., 2004; Kroll et al., 2004).

Use of GIS has streamlined the pre-processing and post-processing of data for environmental modeling; however, common manual methods of GIS are often tedious, time-consuming, and problematic. Repetitive manual inputs can distract a user's concentration, which may result in human errors. Unlike computer programs that have a traceable path, manual GIS operations may be hard to reproduce. To overcome these problems, a GIS toolset has been developed to automate the computation of descriptive statistics for large amounts of raster data over areas of interest.

2. Methodology

ArcGIS 9.0 with its Spatial Analyst extension was used as an application platform. GIS processes were customized using Visual Basic for Applications (VBA) and ArcObjects (Cameron, 2001; Chang, 2005). Fig. 1 provides a schematic diagram of the GIS toolset to calculate descriptive statistics with multiple raster data. Formats for data in this figure are represented with different symbols, and are denoted at the bottom of the figure. Double-lined rectangles represent the two tools developed: Batch Output Table Creation and Batch Descriptive Statistic Calculation.

Descriptive statistics are derived by overlaying raster data representing environmental data onto each area of interest defined by polygons, and computing statistics of values assigned in grids that fall within each area. Calculated statistics are stored in an output table along with the identification (ID) of corresponding polygons. These procedures involve a variety of GIS operations. Moreover, to process a large amount of raster data with many polygons, these operations need to be repeatedly performed. With the developed GIS toolset, these processes can be automated in a batch process. The Batch Output Table Creation tool initializes output tables, and the Batch Descriptive Statistic Calculation tool fills the tables with the calculated statistics. The function of each tool along with its input and output data are explained next.

2.1. Batch output table creation tool

The Batch Output Table Creation tool allows users to create multiple output tables in a batch process. The created output tables may be used to separately store calculated descriptive statistics for different types of raster datasets. An inputted parameter file specifies the physical paths of the Windows file system where the output tables are to be created. An example of this parameter file is found in Fig. 2. In the parameter file, lines starting with a hash sign (#) are comment lines, the tablePath indicates the location of the table to be created, and the tableName indicates the name of the table. Inputted polygon data can include

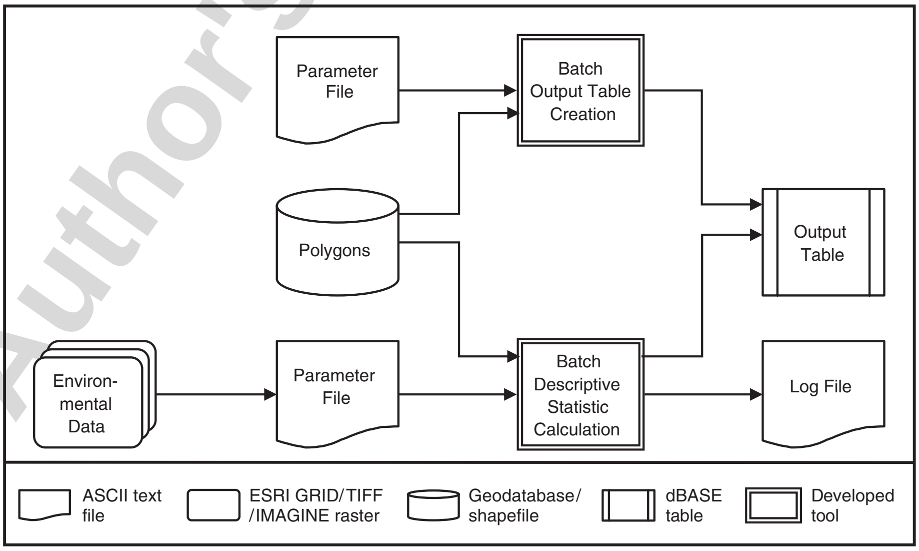
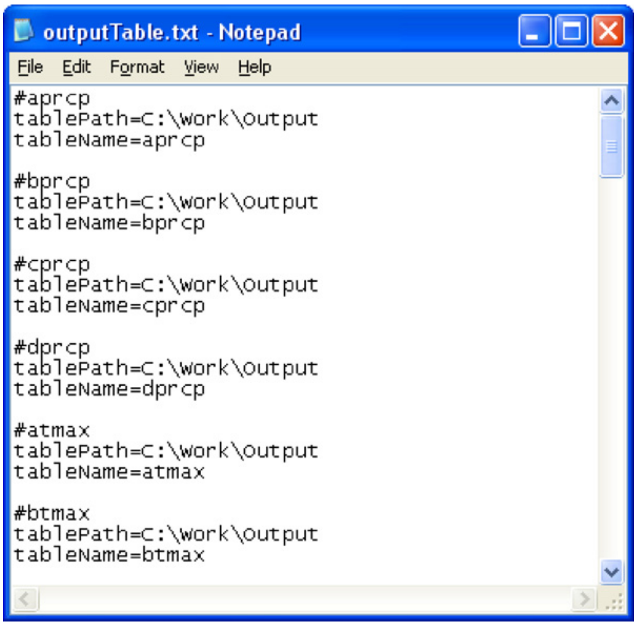


Fig. 1. Schematic diagram of processes performed with developed GIS toolset. Double-lined rectangles represent tools developed in this study.



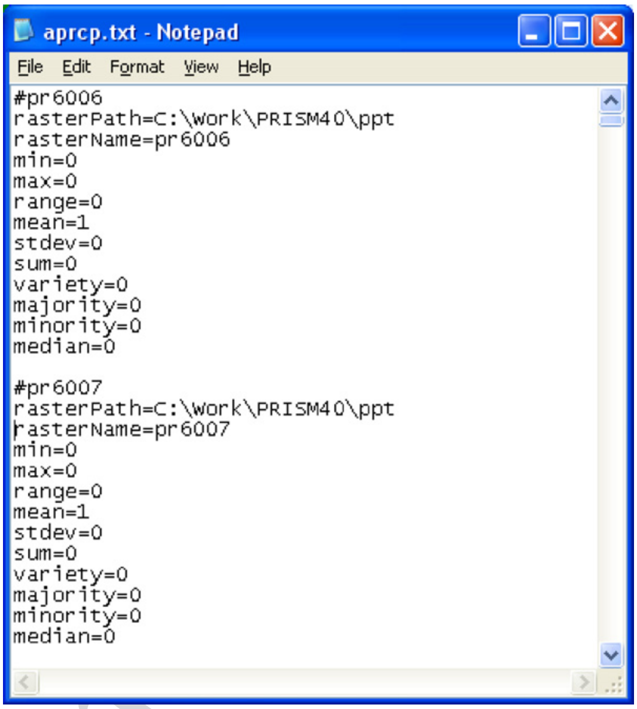
```
#aprcp
tablePath=C:\work\output
tableName=aprcp

#bprcp
tablePath=C:\work\output
tableName=bprcp

#cprcp
tablePath=C:\work\output
tableName=cprcp

#dprcp
tablePath=C:\work\output
tableName=dprcp

#atmax
tablePath=C:\work\output
tableName=atmax

#btmax
tablePath=C:\work\output
tableName=btmax
```

Fig. 2. Example of a Batch Output Table Creation parameter file. Parameter file defines physical paths to output tables to be created.

multiple regions of interest, and can represent any spatial feature such as property, city, state or watershed boundaries. The IDs for each polygon is assigned to a row in the created output tables. The created output tables are in a dBASE format, which is readable by spreadsheet packages, such as MS-Excel.

2.2. *Batch descriptive statistic calculation tool*

The Batch Descriptive Statistic Calculation tool provides an efficient way to calculate descriptive statistics with a large amount of raster data. Physical paths of the Windows file system to multiple raster data and the types of statistics to be computed are specified in a parameter file. Fig. 3 shows an example of such a parameter file. The Batch Descriptive Statistic Calculation tool interprets lines starting with a hash sign (#) as comments and other lines as parameters. Parameters are specified with a predefined keyword, an equal sign (=), and a parameter value. The physical path to the Windows file system folder is defined with the rasterPath keyword, while the name of the raster data is defined with the rasterName keyword. Statistic types are defined with ten keywords: min, max, range, mean, stdev, sum, variety, majority, minority, median, and a parameter value of 0 or 1. A value of 1 indicates the corresponding statistic is calculated and stored, while a value of 0 indicates it is not. The user indicates what statistics should be calculated for each raster data in the parameter file.



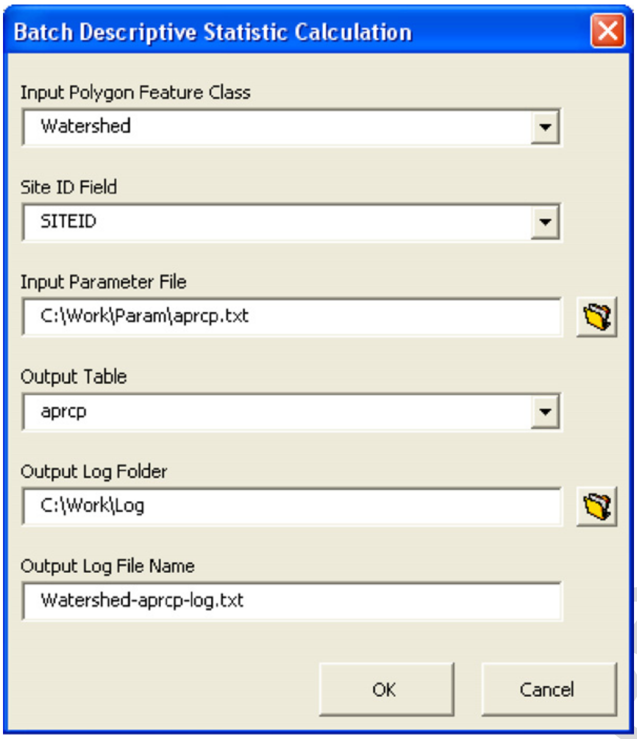
```
#pr6006
rasterPath=C:\work\PRISM40\ppt
rasterName=pr6006
min=0
max=0
range=0
mean=1
stdev=0
sum=0
variety=0
majority=0
minority=0
median=0

#pr6007
rasterPath=C:\work\PRISM40\ppt
rasterName=pr6007
min=0
max=0
range=0
mean=1
stdev=0
sum=0
variety=0
majority=0
minority=0
median=0
```

Fig. 3. Example of a Batch Descriptive Statistic Calculation parameter file. Parameter file defines physical paths to folder and raster data as well as descriptive statistic types to be computed.

The Batch Descriptive Statistic Calculation tool has a graphical user interface that is presented in Fig. 4. With this interface the polygon data, the field of the polygon ID, and the parameter file, output table, and output log name are identified. Reading the parameter file, the Batch Descriptive Statistic Calculation tool overlays each raster data onto the polygons, calculates designated statistics for each polygon, and stores them in the output table. Attribute names for the calculated statistics stored in the output table are automatically created by combining up to the first six letters of the raster data name and three letters indicating the type of descriptive statistic. In addition, to confirm the state of the process after batch processing is completed, a log file is created. The log file records processed raster data paths and names, and error or success information from the batch processing.

3. Case study

A case study is presented in this section, in which a watershed characteristic database was created using the developed GIS toolset. An objective of the case study is to examine the efficiency of the developed GIS toolset in constructing a watershed characteristic database containing data from a related low streamflow prediction study (Kroll et al., 2004).



Input Polygon Feature Class: Watershed
Site ID Field: SITEID
Input Parameter File: C:\Work\Param\aprcp.txt
Output Table: aprcp
Output Log Folder: C:\Work\Log
Output Log File Name: Watershed-aprcp-log.txt

Fig. 4. Batch Descriptive Statistic Calculation tool user interface. Users can specify inputs and outputs for Batch Descriptive Statistic Calculation tool.

3.1. Data employed

Table 1 presents data employed in this study. Polygon data represents watersheds for 35 US Geological Survey (USGS) gauging sites. Each watershed has an ID (i.e. SITEID) as labeled in Fig. 5. A total of 1466 raster datasets representing topography, soil, and climate were employed. The topography data is a Digital Elevation Model (DEM) obtained from the USGS National Elevation Dataset (NED), the soil classification data is from the United States Department of Agriculture's State Soil Geographic (STATSGO) database, while the climate data is from the Spatial Climate Analysis Service's Parameter-elevation Regressions on Independent Slopes Model (PRISM) (Daly et al., 1993). The climate data represents two data sets. The first dataset is spatially distributed average monthly and average annual precipitation for 30 years (1961–90). The second dataset is spatially distributed monthly precipitation, maximum temperature, and minimum temperature for 40 years (1960–99) monthly time series.

3.2. Batch processing

Table 2 provides details of the processed raster datasets, calculated statistics, parameter files, and output tables. A total of 15 output tables were created in a batch process using the developed Batch Output Table Creation tool. Note that one large table could have been created; more tables were created to better organize the output. These tables

Table 1
Data employed

Data type	Data	Resolution	Number of raster data	Data source
Polygon	Watersheds	n/a	n/a	USGS national water information system[a]
Raster	Topography	~30 m	1	USGS national elevation dataset (NED) DEM[b]
	Soil	~1 km	12	USDA state soil geographic (STATSGO)[c]
	Climate	~4 km	12	PRISM 30-year average monthly precipitation[d]
		~4 km	1	PRISM 30-year average annual precipitation[d]
		~4 km	480	PRISM 40-year monthly precipitation[d]
		~4 km	480	PRISM 40-year monthly maximum temperature[d]
		~4 km	480	PRISM 40-year monthly minimum temperature[d]

[a] United States Geological Survey (USGS) NWIS (National Water Information System) Web Data for the Nation 〈http://waterdata.usgs.gov/nwis〉.

[b] USGS, National Elevation Dataset 〈http://ned.usgs.gov/〉.

[c] United States Department of Agriculture (USDA), Natural Resources Conservation Service, National Cartography and Geospatial Center, State Soil Geographic (STATSGO) Database 〈http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo/〉.

[d] Spatial Climate Analysis Service (SCAS), Oregon Climate Service (OCS), PRISM Products 〈http://www.ocs.orst.edu/prism/index.phtml〉.

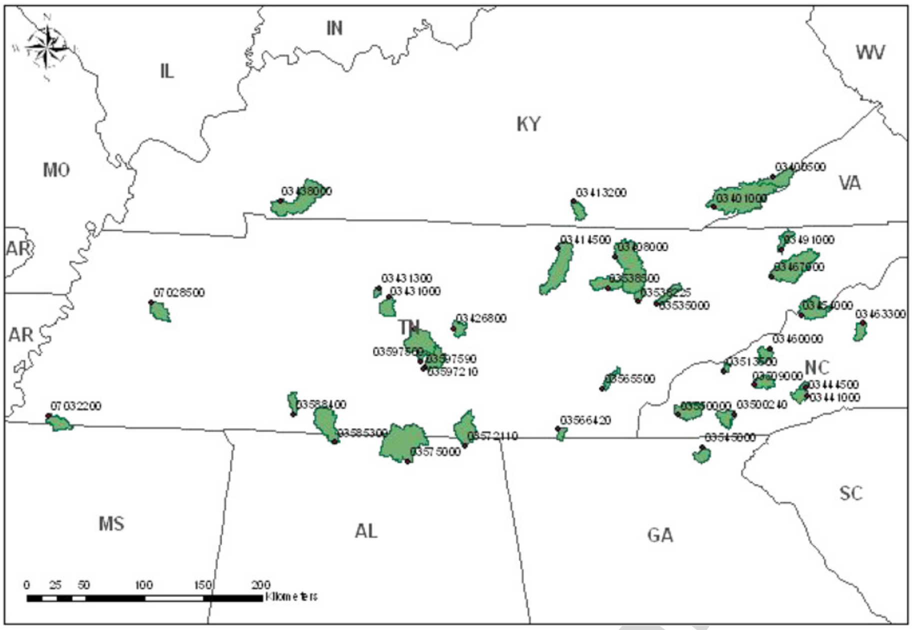


Fig. 5. Watershed boundaries. Thirty-five watershed boundaries are represented by polygons. Eight-digit numbers represent ID (SITEID) for each watershed.

were first initialized with the polygon IDs using the Batch Output Table Creation tool.

For each created output table, a parameter file specifying the raster data and statistic type were prepared. Upon completion of the output tables and parameter files, a batch process of the descriptive calculation was performed. The Batch Descriptive Statistic Calculation tool reads the parameter file and obtains raster data from the defined paths. The raster data was then overlain on 35 polygons (watersheds) and descriptive statistics were computed over each polygon. For each combination of raster data and statistic, an attribute name was assigned. For instance, the mean value calculated for the raster data called PR6006 was assigned an attribute name, PR6006_MEA as illustrated in the output table shown in Fig. 6. Computed statistics were then stored in a cell with the corresponding ID and attribute name in the output table. For all raster data listed in the parameter file, these processes were repeated automatically. In addition to the output tables, a log file is also created that indicates any errors within the batch processing.

4. Results and discussions

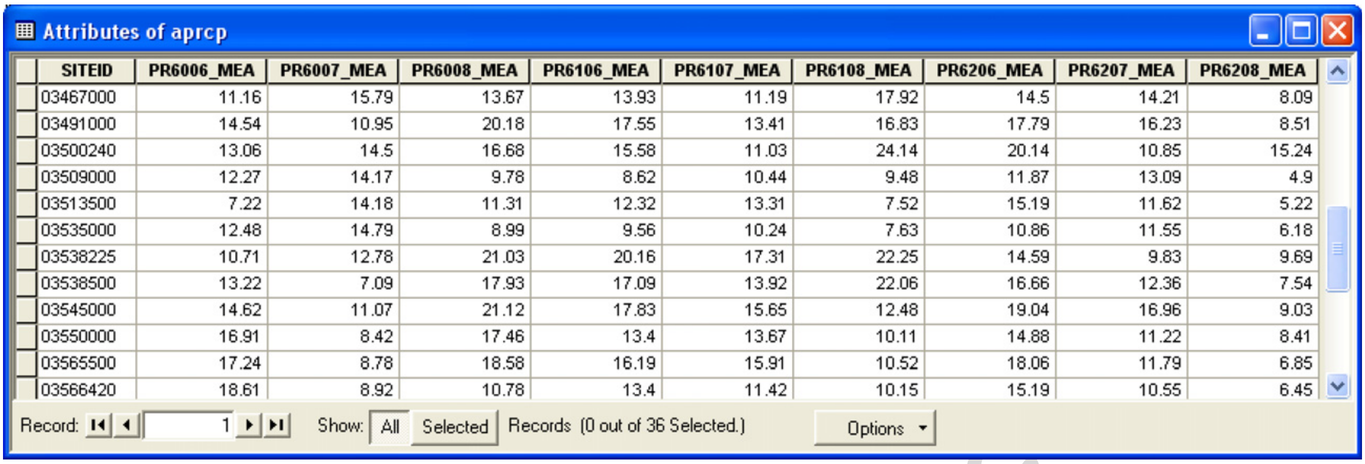
Application of the developed GIS tools to the case study indicates that the batch processing of statistics calculation greatly reduced time and effort to process a large amount of raster data. The required time for batch processes to calculate descriptive statistics with 1466 raster datasets was estimated to be approximately one hour (with a Pentium® 4 CPU of 2.80 GHz and 1.00 GB RAM), while manual operation for the same calculation is estimated to take at least 20 h. Based on these values, it is estimated that the developed statistic processing toolset saved at least 95% of the labor time for the creation of this database.

Manual operation for watershed characteristic extraction would involve specifying raster data, running the Zonal Statistics command of ArcGIS to calculate descriptive statistics, and copying the calculated statistics to an output table. To process a large amount of raster data, these manual operations need to be performed repeatedly, which may potentially produce human errors. Due to a lack of traceability for those manual operations, errors are difficult to detect. On the other hand, the developed GIS toolset has a capability to trace data and procedures. For the developed GIS toolset, sources for human error reside only in the creation of the parameter files and inputs to the user interface. Those possible errors can be detected in the process logging and attribute naming. In addition, if new raster data become available, statistics can be easily calculated with minimal

Table 2
Prepared parameter files and output tables

Data type	Raster data	Number of files	Computed statistics	Raster data description	Parameter file	Output table
Topography	DEM	1	Minimum	Gauge elevation	topo.txt	topo.dbf
			Mean	Mean elevation		
			Range	Watershed relief		
Soil	awch	1	Mean	Maximum available water capacity	soil.txt	soil.dbf
	awcl	1	Mean	Minimum available water capacity		
	bdh	1	Mean	Maximum moist bulk density		
	bdl	1	Mean	Minimum moist bulk density		
	omh	1	Mean	Maximum organic matter content		
	oml	1	Mean	Minimum organic matter content		
	ph	1	Mean	Maximum permeability rate		
	pl	1	Mean	Minimum permeability rate		
	rdh	1	Mean	Maximum bedrock depth		
	rdl	1	Mean	Minimum bedrock depth		
	wdh	1	Mean	Maximum water table depth		
	wdl	1	Mean	Minimum water table depth		
40-year monthly precipitation	prYY[a]06,07,08	120	Mean	June, July, August	aprcp.txt	aprcp.dbf
	prYY[a]09,10,11	120	Mean	September, October, November	bprcp.txt	bprcp.dbf
	prYY[a]12,01,02,03	160	Mean	December, January, February, March	cprcp.txt	cprcp.dbf
	prYY[a]04,05	80	Mean	April, May	dprcp.txt	dprcp.dbf
40-year monthly maximum temperature	maYY[a]06,07,08	120	Mean	June, July, August	atmax.txt	atmax.dbf
	maYY[a]09,10,11	120	Mean	September, October, November	btmax.txt	btmax.dbf
	maYY[a]12,01,02,03	160	Mean	December, January, February, March	ctmax.txt	ctmax.dbf
	maYY[a]04.05	80	Mean	April, May	dtmax.txt	dtmax.dbf
40-year monthly minimum temperature	miYY[a]06,07,08	120	Mean	June, July, August	atmin.txt	atmin.dbf
	miYY[a]09,10,11	120	Mean	September, October, November	btmin.txt	btmin.dbf
	miYY[a]12,01,02,03	160	Mean	December, January, February, March	ctmin.txt	ctmin.dbf
	miYY[a]04,05	80	Mean	April, May	dtmin.txt	dtmin.dbf
30-year average monthly and annual precipitation	pt01	1	Mean	January monthly average	ppt30.txt	ppt30.dbf
	pt02	1	Mean	February monthly average		
	pt03	1	Mean	March monthly average		
	pt04	1	Mean	April monthly average		
	pt05	1	Mean	May monthly average		
	pt06	1	Mean	June monthly average		
	pt07	1	Mean	July monthly average		
	pt08	1	Mean	August monthly average		
	pt09	1	Mean	September monthly average		
	pt10	1	Mean	October monthly average		
	pt11	1	Mean	November monthly average		
	pt12	1	Mean	December monthly average		
	ptan	1	Mean	Annual average		

[a] YY represents lower two digits (60–99) of 40 years (1960–99).



SITEID	PR6006_MEA	PR6007_MEA	PR6008_MEA	PR6106_MEA	PR6107_MEA	PR6108_MEA	PR6206_MEA	PR6207_MEA	PR6208_MEA
03467000	11.16	15.79	13.67	13.93	11.19	17.92	14.5	14.21	8.09
03491000	14.54	10.95	20.18	17.55	13.41	16.83	17.79	16.23	8.51
03500240	13.06	14.5	16.68	15.58	11.03	24.14	20.14	10.85	15.24
03509000	12.27	14.17	9.78	8.62	10.44	9.48	11.87	13.09	4.9
03513500	7.22	14.18	11.31	12.32	13.31	7.52	15.19	11.62	5.22
03535000	12.48	14.79	8.99	9.56	10.24	7.63	10.86	11.55	6.18
03538225	10.71	12.78	21.03	20.16	17.31	22.25	14.59	9.83	9.69
03538500	13.22	7.09	17.93	17.09	13.92	22.06	16.66	12.36	7.54
03545000	14.62	11.07	21.12	17.83	15.65	12.48	19.04	16.96	9.03
03550000	16.91	8.42	17.46	13.4	13.67	10.11	14.88	11.22	8.41
03565500	17.24	8.78	18.58	16.19	15.91	10.52	18.06	11.79	6.85
03566420	18.61	8.92	10.78	13.4	11.42	10.15	15.19	10.55	6.45

Records (0 out of 36 Selected.)

Fig. 6. Example of an output table with computed values stored. Computed statistics for each watershed are stored in an output table. Attribute names for computed statistics are defined from raster data name used and descriptive statistic computed.

changes to the parameter files. Procedures implemented in the developed tools can also be easily transferred to other study areas or other environmental studies.

5. Conclusions

A GIS toolset has been developed in an ArcGIS 9.0 platform. Development was motivated by a need for efficient processing of a large amount of raster data, as well as a lack of generic GIS tools that can be used for a variety of environmental studies. The developed GIS toolset has been utilized for a case study that required the extraction of watershed characteristics for 35 watersheds from approximately 1500 raster datasets. Without this GIS toolset, manually deriving watershed characteristics with such a large amount of raster data would be tedious, time-consuming, and prone to errors. A case study indicates that the developed GIS toolset greatly reduced time and effort to calculate descriptive statistics across regions of interest with a large amount of raster data. The developed GIS toolset improved the traceability, reproducibility, and transferability of GIS procedures. State, county, and town boundaries, as well as any property boundaries that can be represented by polygons, can be an input to the developed toolset, and any number of raster datasets can be processed by the automated batch process. The developed GIS toolset is versatile and can aid in a wide variety of environmental studies, and is available for free download at www.esf.edu/erfeg/kroll.

Acknowledgements

The authors would like to acknowledge the US Geological Survey State Water Resources Research Institute (WRRI) Program (Grant 2003NY33G) and the USDA Cooperative State Research, Education, and Extension Service (CSREES) Program (Grant NYR-2005-03897) which provided financial assistance to this research. This research has not been subjected to any USGS or USDA review, and therefore does not necessarily reflect the views of those agencies, and no official endorsement should be inferred. The others would also like to thank the two anonymous reviewers who provided insightful comments that greatly improved this manuscript.

References

Bedient, P.B., Hoblit, B.C., Gladwell, D.C., Vieux, B.E., 2000. NEXRAD radar for flood prediction in Houston. Journal of Hydrologic Engineering 5 (3), 269–277.

Cameron, E., 2001. Developing with ArcObjects. In: Zeiler, M. (Ed.), Exploring ArcObjects, vol. 1—Applications and Cartography. ESRI Press, Redlands, CA, p. 725.

Chang, K., 2005. Programming ArcObjects with VBA A Task-Oriented Approach. CRC Press, Boca Raton, FL (p. 352).

Daly, C., Neilson, R.P., Phillips, D.L., 1993. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33, 140–158.

He, C., 2003. Integration of geographic information systems and simulation model for watershed management. Environmental Modeling and Software 18, 809–813.

Kroll, C.N., Luz, J.G., Allen, T.B., Vogel, R.M., 2004. Developing a watershed characteristics database of improve low streamflow prediction. Journal of Hydrologic Engineering March/April, 116–125.

Maidment, D.R., 1996. Environmental modeling within GIS. In: Goodchild, M.F., et al. (Eds.), GIS and Environmental Modeling: Progress and Research Issues. GIS World Books, Colorado, pp. 315–323.

Neary, V.S., Habib, E., Fleming, M., 2004. Hydrologic modeling with NEXRAD precipitation in middle Tennessee. Journal of Hydrologic Engineering September/October, 339–349.

Pullar, D., Springer, D., 2000. Towards integrating GIS and catchment models. Environmental Modeling and Software 15, 451–459.

Rao, M.N., Waits, D.A., Neilsen, M.L., 2000. A GIS-based modeling approach for implementation of sustainable farm management practices. Environmental Modeling and Software 15, 745–753.