Python Script for Precipitation Statistics

Jennifer McCollum

Programming for Geospatial Science GISC 3200K Fall 2020

University of North Georgia, IESA

ABSTRACT

The purpose of this project was to develop an automated extraction process of precipitation data using Python programming for the *Programming for Geospatial Science* course taught at the University of North Georgia (UNG). I chose this project because the UNG Environmental Protection Agency (EPA) research team is currently studying over 30 years of water quality sample data from the UNG Water Lab in Dahlonega. Their goal was to correlate total phosphorous with land use and stormwater runoff. Initial tests confirmed an issue where precipitation data skewed the results because of extreme outliers. A tool was needed to extract and sort the precipitation data to see total rainfall, average rainfall, and to find the highest and lowest rainfall months. The idea was to extract and filter days that do not have precipitation and to find extreme precipitation occurrences. Python was chosen for this project because it has become the primary Geographical Information Systems (GIS) programming language for ArcGIS (Etherington, 2015). Thomas Etherington, at the Institute for Applied Ecology New Zealand, School of Applied Sciences, Auckland University of Technology believes that by teaching his students Python, they will be able to automate tasks, modify analytical processes, and more importantly, "asks questions that could not be asked using pre-built software" (2015). The focus of this paper introduces why I developed a GIS-based Python program for automated extraction of precipitation data and the material and methods used in the process.

Keywords: Python, Coding, Data analysis

UNG
UNIVERSITY of NORTH GEORGIA™
LEWIS F. ROGERS INSTITUTE FOR ENVIRONMENTAL AND SPATIAL ANALYSIS

Submitted to Dr. Huidae Cho *December 4, 2020*

INTRODUCTION

Background

Lake Lanier is a reservoir in North Georgia that provides drinking water to several large cities in the area. To comply with water quality standards, the Georgia Environmental Protection Division conducts a Total Maximum Daily Load (TMDL) evaluation (EPD, 2017). The TMDL shows the relationship between sources of pollutants and water quality conditions that include several parameters (EPD, 2017). One of those parameters is precipitation data. Precipitation data can be used to calculate runoff and to see how pollutants are transported to larger bodies of water. For example, during a large rainfall event, stormwater runoff can carry excess nutrients from urban and agriculture areas into large bodies of water, like Lake Lanier. According to Hun-Kyun Bae from the Department of Global Environment Keimyung University, increased rainfall events can affect water quality samples that are collected over time (2013). Bae performed a study that analyzed the effects of rainfall events on water quality, and their results showed that certain amounts of rainfall and the frequency of rainfall do affect the water quality.

The UNG Water Lab started collecting the samples around Lake Lanier in 1987 to create a continuous baseline water quality-monitoring program of the Lake Lanier Watershed (UNG, 2020). The samples are collected from eleven sites in the Upper Chattahoochee River Basin from 1987 to 2020 (UNG, 2020). Researchers from the UNG Chemistry, Biochemistry, and the Environmental Spatial Analysis Departments used the water quality data to analyze total phosphorus (P). Their first goal in working with this data was to see where the highest levels of total phosphorus concentration occurred. Next, they correlated the Total P with land use and stormwater runoff. However, during an initial run of the water quality data, there was an issue with the water quality being skewed by precipitation measurement outliers. Therefore, a group decision was made to try and sort the precipitation data in a way that allowed them to see if the non-rainy days compared to the rainy days affected the outcome of the water quality samples.

Precipitation Analysis

Precipitation data is key to environmental modeling. It is useful for hydrology, water quality, climate, erosion, and agricultural research (Sitterson, et al., 2020). However, precipitation data can be complicated and sometimes tricky to decipher. Because of this, researchers at the United States Environmental Protection Agency (EPA) Office of Research and Development National Exposure Research Laboratory in Athens Georgia performed a study to find the benefit on using web services tools to incorporate automatic retrieval and comparison of precipitation data (Sitterson, et al., 2020). The researchers determined that finding the data and preprocessing it can be time-consuming, challenging, and could have a "major implication on environmental modeling results" (Sitterson, et al., 2020). Their solution involved using a tool called Hydrological Micro Services Precipitation Comparison and Analysis Tool (HMS-PCAT) which gives users access to multiple precipitation datasets at the same time. It pulls the data from the web and combines them into one file for side by side comparison. The benefit of using this program allows scientists who are not skilled programmers to quickly access data (Sitterson, et al., 2020).

Another study points out how essential precipitation data is, and also how daunting it can be to navigate through the process of finding quality data. Sean Zeiger from the Institute of Water Security, and Jason Hubbart from the Davis College, Schools of Agriculture and Food, and Natural Resources at West Virginia University, performed a study to minimize precipitation model uncertainty in the Soil and Water Assessment Tool (SWAT) by using multiple methods to find the mean areal precipitation (MAP) estimates (2017). They suggest that there are very few publications on precipitation data and how it affects the SWAT model output (Zeiger and Hubbart, 2017). Because, not only can topography, land use, or rain gauge miscalculations occur, it is critical to examine the quality of data before performing the analysis (Zieger and Hubbart, 2017). After 19 methods were used to find MAP estimates, Zeiger and Hubbart found that inverse-distance weighted, linear polynomial interpolation or multiquadric function methods were the best ways to improve SWAT model simulations (2017).

While precipitation data can be complicated and daunting sometimes, it can also unveil secrets that might otherwise be hidden. According to Constantinos Doskas, head of the IT and Security Department of Olympus, he believes it is the analyst's responsibility to find abstracted facts inside data to provide answers to questions (2020). Doskas recently performed a study that focuses on how to "organize data in logical groups" using python programming which made it easy for researchers to understand and use (2020).

Python

The purpose of this project was to develop a Python program that extracts rainfall data statistics for future analysis. This Python programming allows researchers to extract and filter months with extreme precipitation occurrences. Python was chosen because it has many benefits when working with geospatial data. In fact, GIS programming is becoming more popular in geographical disciplines and is deemed almost necessary for students to effectively market themselves to future employers (Gallagher and Trendafilov, 2018). Michael Gallagher, from St. Bonaventure University, and Rossen Trendafilov from St. Thomas Aquinas College framed a study around the fact that Python is easier to code and debug (2018). Their study called *R Vs. Python: Ease of Use and Numerical Accuracy*, compares the programming languages R and Python. They were determined to find out why students were drawn to Python, hinting that it might be because of how trendy Python is and not so much about how easy it is (Gallagher and Trendafilov, 2018). However, their conclusion whole-heartedly states that Python is easy to use and made for scripting and automating processes, however those that are interested in larger data should learn R as a coding language which is better to handle statistical and analytical analysis (Gallagher and Trendafilov, 2018).

Using Python to automate any process will save time in the long run which was proven by Ashton Greer, Zachary Wilbanks, Leah Clifton, Bradford Wilson, and Andrew Graettinger from the Department of Civil, Construction, and Environmental Engineering at the University of Alabama (2018). They designed a Python module that automatically designs a culvert "within a

single-computational platform" (Greer et al., 2018). The results were produced in minutes, rather than hours.

Study Area

Figure 1 below shows the study area of Lake Lanier watershed located in North Georgia. Lake Lanier was built and is operated by the U.S. Army Corps of Engineers. They maintain flood control and water supply. The lake is visited by millions if people every year. The lake itself is close to 59 square miles of water (Wikipedia 2020).

Figure 1. *Study area.*

MATERIAL

Precipitation Data

Daily precipitation for the City of Gainesville were obtained from the AN81d PRISM dataset. This dataset spans January 1, 2016 to December 31, 2016, one full year. You can access this dataset from the PRISM website: https://prism.oregonstate.edu/explorer/. It covers one grid, 4km spatial resolution and elevation at 1165 ft. By choosing the Interpolate option, an inverse-distance squared weighting is applied, and the surrounding grid cell centers will be added (PRISM, 2020). A study was done in 2017 by Christopher Daly, et al., called *High-Resolution Precipitation Mapping in a Mountainous Watershed: Ground Truth for Evaluating Uncertainty in a National Precipitation Dataset*, that showed how difficult it was to evaluate potential sources of uncertainty in the interpolated PRISM grids because accurate precipitation data is usually unknown (2017). In order to accommodate this uncertainty, a cross-validation must be performed (Daly, et al., 2017).

Since this study is only looking at amount of rainfall for one point, the interpolate option was not selected, however that is something that can be looked at in the future. According to EPA's Office of Research and Development, PRISM has one of the highest correlation coefficients between 1981 to 2017 regarding variability of precipitation datasets in diverse environments in 17 different regions (Sitterson, et al., 2020). Figure 2 shows selected attributes for the precipitation data selected for this project on the PRISM website.

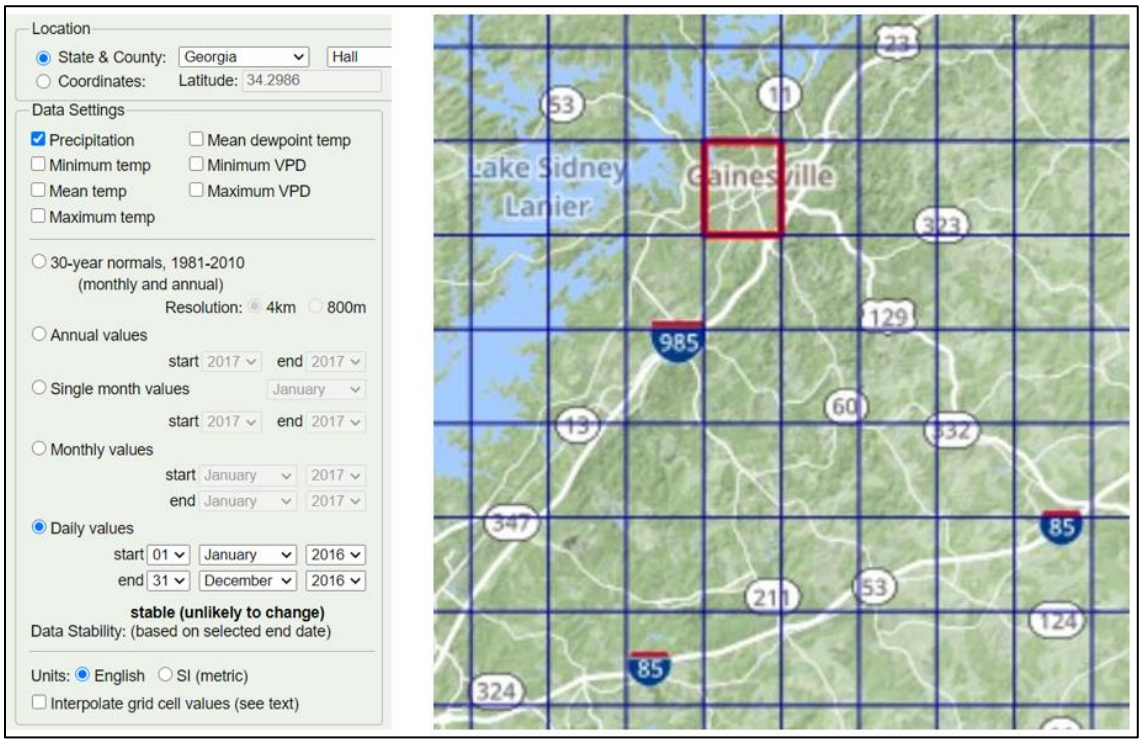

Figure 2. Time series values for individual locations. https://prism.oregonstate.edu/explorer/

One of the benefits of the PRISM website is that it gives an overview of precipitation for your selected time period. In the figure below, it shows which month has the most rainfall in 2016.

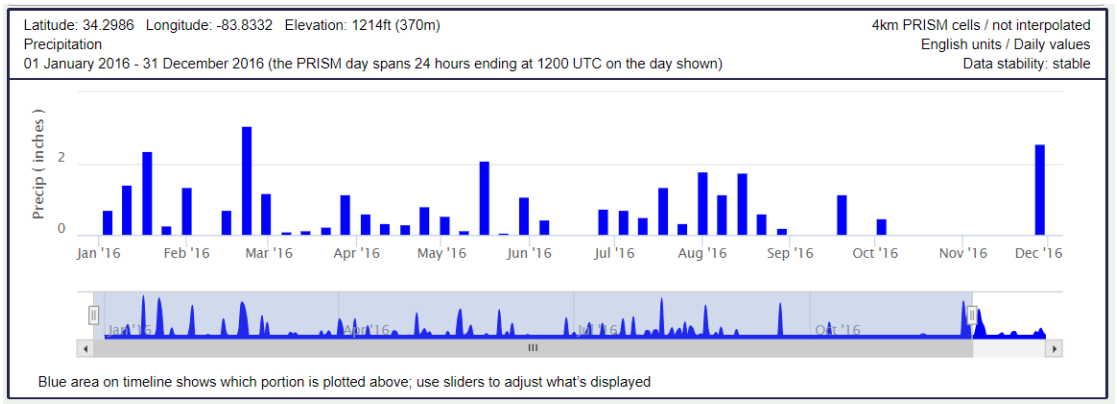

Figure 3. PRISM graph shows precipitation timeline.

METHOD

Precipitation Data Setup

Before the python script can process correctly, the precipitation data needs to be set up properly to function. Excel formulas are used to quickly split up the date into day, month, and year. Another formula is used to turn the month number into a month name for display purposes only. Below are the steps for the precipitation data setup (see figure 5) and the excel formulas used (see figure 4).

Precipitation Data Setup

1. Download .csv data from PRISM
2. Open .csv file
3. Add new columns: “year”, “month”, “day”, “monthNum”, and “siteNum”
4. Drag columns so they are in this order and spelled correctly as shown in figure 5
5. Save as a .csv

Excel Formulas

Split the day	=DAY(D2)
Split the month	=MONTH(D2)
Split the year	=YEAR(D2)
Turn month to name	=TEXT(F2)*29,”mmmmmmmmm”)

Figure 4. Excel formulas used to set up the precipitation data table.

Precipitation Data Setup (.csv file)

year	month	inches_per_day	long	lat	date	day	monthNum	siteNum
2016	January	0	-83.9754	34.4696	1/1/2016	1	1	8
2016	January	0	-83.9754	34.4696	1/2/2016	1	1	8
2016	January	0.2	-83.9754	34.4696	1/3/2016	1	1	8
2016	January	0.09	-83.9754	34.4696	1/4/2016	1	1	8

Figure 5. Sample of precipitation data. Use this as a template for future precipitation data.

Python Scripting

ArcGIS Pro offers multiple geoprocessing tools that are easy to use, however Python scripting is much more efficient (ArcGIS, 2020). The Python window is an interactive Python interpreter that executes Python directly inside ArcGIS Pro. You can create the script in any integrated development environment (IDE), such as ArcGIS Pro Python window, or PyCharm.

DISCUSSION & RESULTS

The goal of the Python script was to extract rainfall statistics and display them in a text file. The Python script was successful. The expected rainfall statistics processed the correct output into a text file (see figure 6).

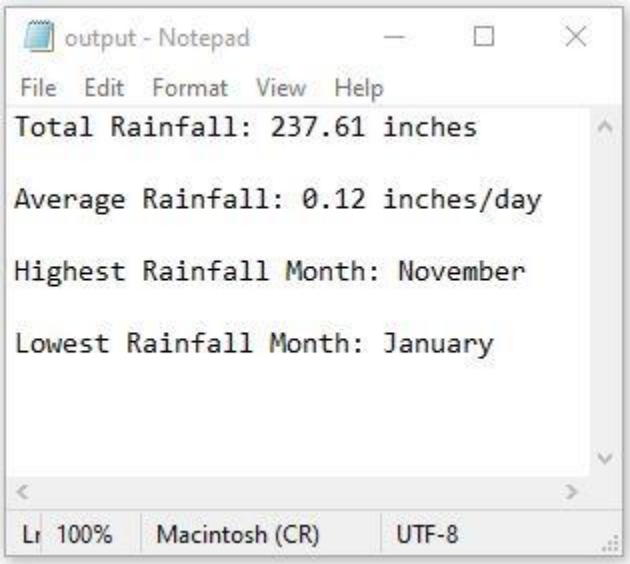

Figure 6. This is an example of the results output text file

Using Python script to execute simple statistics turned out to be relatively easy, yet extremely useful and user friendly. However, changes can be made in the future to enhance this code to be even more beneficial. For instance, this script does not display the data on the map, rather it outputs to a text file. If the script added a rainfall statistics layer to the map, advanced analysis could be performed.

Overall, there are many variables that could affect the outcome of precipitation data analysis. For example, some variables that could cause problems in the future are rain gauge miscalculations, flow rate during a rainstorm, the amount of discharge from upstream, and which area the non-point source pollution comes from. The seasonal cycle might also affect the outcome, as well as dormant seasons. With all the different variables that are possible, being able to automate precipitation data will save time and money.

REFERENCES

Bae, H. (2013). Changes of River's Water Quality Responded to Rainfall Events. *Environment and Ecology Research* 1(1): 21-25. http://www.hrpub.org/download/201307/eer.2013.010103.pdf.

Daly, C., Slater, M., Roberti, J., Laseter, S., Swift, L. 2017. High-Resolution Precipitation Mapping in a Mountainous Watershed: Ground Truth for Evaluating Uncertainty in a National Precipitation Dataset. *Royal Meteorological Society International Journal of Climatology*. Retrieved November 2020. https://prism.oregonstate.edu/documents/pubs/2017IJOC_HiResPrecipMapping_daly_combined.pdf

Devine, J.K. 2018. EPA Awarded $100,000 Grant to Institute for Environmental and Spatial Analysis Faculty. *UNG Newsroom*. Retrieved October 2020. https://ung.edu/news/articles/2018/10/epa-awarded-100,000-grant-to-institute-for-environmental-and-spatial-analysis-faculty.php

Doskas, C. 2020. The Python Programming Language: Relational Databases. Developing and Connecting Cybersecurity Leaders Globally. Retrieved November 2020. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=tsh&AN=143293662&site=eds-live&scope=site&custid=ns235470

EPD. 2017. Final Total Maximum Daily Load Evaluation for Lake Lanier in the Chattahoochee River Basin for Chlorophyll a. *The Georgia Department of Natural Resources Environmental Protection Division Atlanta, Georgia.* Retrieved October 2020. https://epd.georgia.gov/document>publication>download>pdf

ESRI. 2020. Using the Python Window. ESRI ArcGIS for Desktop. Retrieved November 2020. https://desktop.arcgis.com/en/arcmap/10.3/analyze/executing-tools/using-the-python-window.htm

Etherington, T. 2015. Teaching Introductory GIS Programming to Geographers Using an Open Source Python Approach. *Journal of Geography in Higher Education, 2016*. Vol. 40, No. 1, 117-130. Retrieved November 2020. http://dx.doi.org/10.1080/03098265.2015.1086981

Gallagher, M., Trendafilov, R. 2018. R VS. Python: Ease of Use and Numerical Accuracy. *Journal of Business and Accounting*. Retrieved November 2020. http://search.ebscohost.com.proxygsu-nga1.galileo.usg.edu/login.aspx?direct=true&AuthType=ip,shib&db=bth&AN=134412360&site=eds-live&scope=site&custid=ns235470

Greer, A., Wilbanks, Z., Clifton, L., Wilson, B., Graettinger, A. 2018. GIS-Enabled Culvert Design: A Case Study in Tuscaloosa, Alabama. *Advances in Civil Engineering. Vol. 2018. Article ID 4648134, 10 p.* Retrieved November 2020. http://search.ebscohost.com.proxygsu-nga1.galileo.usg.edu/login.aspx?direct=true&AuthType=ip,shib&db=edsdoj&AN=edsdoj.b3697861fc44485870b36c88c719815&site=eds-live&scope=site&custid=ns235470

PRISM. 2020. Time Series Values for Individuals Locations. *PRISM Climate Group Northwest Alliance for Computational Science and Engineering. Retrieved November 2020.* https://prism.oregonstate.edu/explorer/

Sitterson, J., Sinnathamby, S., Parmar, R., Koblich, J., Wolfe, K., & Knightes, C. D. (2020) Demonstration of an Online Web Service Tool Incorporation Automatic Retrieval and Comparison of Precipitation Data. *Environmental Modeling and Software, 123.* Retrieved November 2020. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edselp&AN=S1364815219306279&site=eds-live&scope=site&custid=ns235470

UNG. 2020. Water Lab. *Environmental Leadership Center. Retrieved October 2020.* https://ung.edu/environmental-leadership-center/water-lab/index.php

Zeiger, S., Hubbart, J. (2017). An Assessment of Mean Areal Precipitation Methods on Simulated Stream Flow: A SWAT Model Performance Assessment. *Water* (20734441), 9(7), 459. Retrieved November 2020. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eih&AN=124353333&site=eds-live&scope=site&custid=ns235470