Quantifying Difference of Satellite- and Ground-measured Precipitation Dataset

Data Management Plan

Ankit Ghanghas, and Pin-Ching Li

Lyles School of Civil Engineering, Purdue University

Purpose

This data management plan for Satellite and Ground Comparison Project is focusing on managing the datasets obtained from the open sources provided by NASA and NOAA online. This DMP is built up for a short-term project lasting for at least half a year. Our result will be stored on the cloud and shared with our lab members. Because there is no founder for this project now, the place for storage of datasets would be our personal computer or cloud space. Though there is no sharing platform for people to access our dataset, our dataset is accessible once people contact us personally.

Online Resources: Data Types, Formats, Standards & Capture Methods

Two types of dataset are collected: rainfall data collected by National Oceanic and Atmospheric Administration's (NOAA's), and the Global Precipitation Measurement (GPM) by NASA. GPM is applied as the observation by satellite. GPM dataset can be obtained from the FTP in *NASA precipitation measurement missions* website (https://pmm.nasa.gov/data-access/downloads/gpm). GPM rainfall raster is recorded as .HDF5 file in a 30 minutes time frame. The spatial resolution of GPM is 0.1 degree to 30 min. There are some ways to download GPM files from FTP. Either writing a python script with ftplib, or running our own batch file from Command Prompt could successfully download the files. .HDF5 files can be read in Python. In the United States, rainfall data is available for download at the National Climate Data Center (NCDC). With the script written in Python language, the precipitation datasets from NCDC can be downloaded automatically.

Data and Metadata

Our project will generate csv files for each of the USGS ground weather stations under study. These files will be a time series of half hourly ground observed precipitation measurements and satellite precipitation measurements for the station. This dataset will be encapsulation of already existing precipitation data for certain locations which has been produced using two different methodologies.

We choose to store the analysed results in csv files because they are smaller and efficient to store long series of numerical data (with fewer columns - in our case just 3 namely, Time,satellite precipitation and observed ground precipitation). The dataset will come with a data dictionary (metadata file) providing the latitude and longitude coordinates of the ground stations, format of the data and type of data elements and the naming convention of the output and input files.

Data Sharing.

Since our dataset is encapsulation of already accessible open-source dataset, our data will be available free of cost to whoever is interested in using the dataset. The interested will have to email and ask for data to one of the investigators in the project. The interested should provide reference to the dataset in any

publications or presentations using original or even modified dataset. The user can change the materials to their needs. The dataset will be available for sharing as the project progresses forward.

The user is bound to follow the re-use and re-distribution restrictions and rules for USGS and NASA' GPM datasets. The investigators of this project do not add any further restrictions provided that the reused data cites the original dataset.

Project Regulations and Data Preservation:

All sequence data from this project will be deposited into a one drive cloud folder called "Precip_2020ABE65100". This folder is shared between authors. Both of us can access and edit files in the folder. However, the document worked by both of us, such as the report of results, needs a naming convention for it. It's now assumed to be the filename underscore date underscore author. To avoid the redundancy occurring when we stored the files repeatedly, the old files would be downloaded to our own PC and deleted to release the storage at the last Friday of every month. The backup plan would be stored files in our two ECN machines and weekly backup on an external hard drive. After the project is completed, the data and the codes will be stored on Fortess (Purdue's long term storage system) for long term preservation. Our advisor Prof. Venkatesh Merwade will have access to this long term preserved data for future use and distribution.

The specific naming rule of documents is made. A readme file for any report and code script is demanded. Also, metadata for processed dataset is demanded. People who create the files are responsible for adding these files and comments inside the files. The naming rule of file is to add the date of when the document gets created first. After that, the abbreviation of the author's name is added after the date.

Data Type	Where Deposited	When #	By Whom
raw data from NASA and NOAA	PC and portable hard drive	Anytime	co-author and I
Analysis Result (Graphs, Tables)	One drive, PC and portable hard drive	Within a week when the result is validated by developer	co-author and I
Project Reports	One drive, PC and portable hard drive	One week before the deadline	co-author and I

Code Scripts	One drive, (Github), PC and portable hard drive	Update every week since it is created	co-author and I

Intellectual Property:

Because this project is for research purpose for now, there is no platform or server for sharing to the public. Most of the codes in this project are kept private until the work of publication is done. However the dataset will be available for use. The project data and codes will be handed over to our advisor Prof. Merwade upon the completion of the project for future use and long term usability.