Design Document for Global Surface Water Research Project

Dataset:

Global Surface Water open-source Datasets:
https://global-surface-water.appspot.com

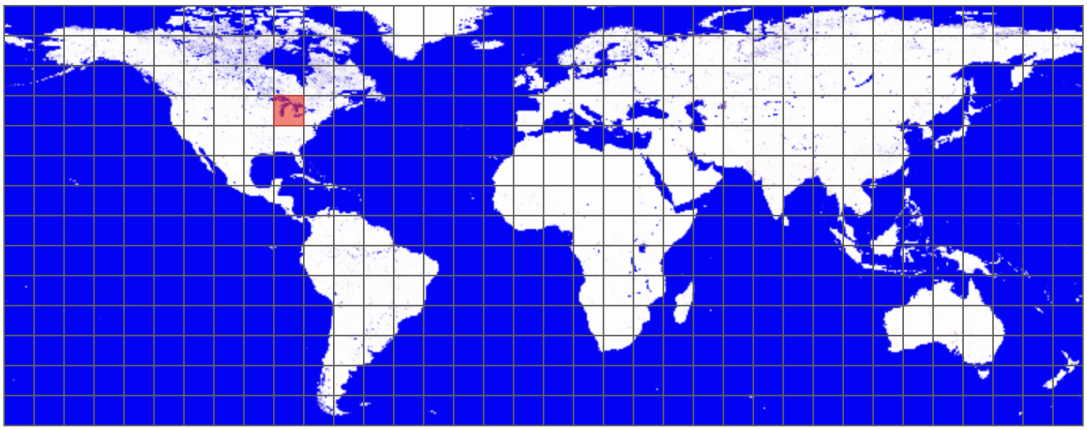
Datasets downloads:
https://global-surface-water.appspot.com/download

In the download page, we have many options for downloading datasets.

We can download sections of the full dataset by tiles (10deg x 10deg of world map) for each attribute (occurrence, change, seasonality…). Below is the illustration of selecting one tile from the map.

Individual 10°x10° files

The Global Surface Water data are available to download in tiles 10°x10° from the map shown below. Click on the tile to show a list of the available datasets. Each one of these datasets is a hyperlink to the *.tif file.



Granule with top-left corner at 90W, 50N:

Occurrence: https://storage.googleapis.com/global-surface-water/downloads2020/occurrence/occurrence_90W_50Nv1_3_2020.tif
Change: https://storage.googleapis.com/global-surface-water/downloads2020/change/change_90W_50Nv1_3_2020.tif
Seasonality 2020: https://storage.googleapis.com/global-surface-water/downloads2020/seasonality/seasonality_90W_50Nv1_3_2020.tif
Recurrence: https://storage.googleapis.com/global-surface-water/downloads2020/recurrence/recurrence_90W_50Nv1_3_2020.tif
Transitions: https://storage.googleapis.com/global-surface-water/downloads2020/transitions/transitions_90W_50Nv1_3_2020.tif
Maximum extent: https://storage.googleapis.com/global-surface-water/downloads2020/extent/extent_90W_50Nv1_3_2020.tif

We can also download the full dataset with the provided instructions on the page.

Datasets are stored as TIF files, each tile consists of 40,000 x 40,000 pixels and individual pixels are valued between 0-255. Values are interpreted differently in different attributes of the data, detailed information on the data format can be found on the github page.

https://github.com/Tim041/Global-Surface-Water-Research

Initial Data Processing:

From the previous section, we see that for each tile, the data is stored in a TIF file with 40,000 x 40,000 pixels. The file size for each tile attribute is about 100MB on average, there are 504 tiles in this map and each tile has 8 different attributes which add up to around 400GB. Thus, we need to find a more effective representation for those dataset.

For each of the tiles, we currently set the final output to be a 10 x 10 CSV file for future ML usage. For each tile, the script will determine the attributes and apply rules to accommodate the data points compression accordingly. For this project, we divided the original 40,000 x 40,000 pixels into 100 sub groups, each containing 4000 x 4000 pixels. For each of the 100 sub groups, the average number will be reported and stored as the compressed data, and the result will be stored into a CSV file.

The input TIF datasets are set to be stored in the Data folder, and the final CSV files will be generated in the Result folder.