# Optimizing Hydroelectric Station Locations Using Data Analytics

**Steven Xie**
New York University
srx201@nyu.edu

**Koji Liu**
New York University
ml7324@nyu.edu

**Vishesh Goyal**
New York University
vkg9435@nyu.edu

## Abstract

This paper aims to identify the optimal locations for siting hydroelectric stations. The pipeline starts with gathering data from three key blocks: Environmental Variables, Infrastructure Constraints, and Socio-Economic Factors. We then integrate data from these factors to analyze potential locations. Our approach considers the best sites by not only technical feasibility but also out of social and economical considerations. We evaluate the results by calculation of scores and rankings. The results agrees with our predictions of the significant elements in determining siting hydropower. We hope this paper serves as a guide for strategic hydropower development and underscore the importance of site selection in energy infrastructure projects.

**Keywords: Hydroelectric power siting; Socio-Economic Factors in Energy Planning; Energy Efficiency Analysis; GIS; Elevation**

## 1 Introduction

In recent years, the movement for switching from non renewable and environmentally polluting energy sources such as coal and oil towards cleaner, more renewable, and more efficient energy has been gaining traction in large amounts worldwide. Among the contenders for new energy sources, hydroelectric power generation sits near the top of the list; hydroelectric power creates minimal pollution, is completely renewable due to the water cycle, and can provide huge quantities of power [Zaidi and Khan (2018)]. But there important considerations when constructing a hydroelectric power station. Volume of water [Wei, Li, et al. (2020)], water flow [Huang and Yan (2009)], and distance of electrical power transfer all make the location of a hydroelectric power station incredibly important to the power the station can produce for consumers. Our goal in this study is to analyze possible locations for hydroelectric power stations, rank the locations by best energy output and efficiency, and ultimately determine the best locations for new hydroelectric power stations. We will first start by gathering data from 3 different datasets including precipitation, topography, and population. After formatting them for upload to HDFS, we will use MapReduce to clean the data and transform it into an easily usable form while using spark to perform EDAs on the datasets. Then the cleaned datasets will be merged and we will use a scoring function to rank possible locations enabling us to visualize these rankings as a heat map and perform factor analysis on our datasets.

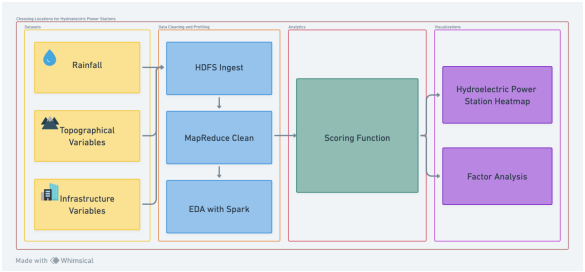


Figure 1: Data Flowchart

## 2 Motivation

Renewable energy has been a staple of energy production for the past century following multiple energy crises and the looming threat of climate change. Hydroelectric power is not only reliable, but also economical and safe. However, one of the biggest concerns with hydroelectricity has always been a lack of suitable places to construct these hydroelectricity generators. Drought/climate irregularities, ecological impact, locality, and topography are among the many factors that must be considered to launch a hydroelectricity project. (Askari et al., 2015) The bulk of cost for hydroelectric generation is in the initial expense of building a dam and generators. This is why we believe that the location of hydroelectricity stations plays a vital

role in renewable energy.

## 3 Related Work

### 3.1 Impact of elevation on hydroelectric station siting

There are several studies that supports the potential effects of elevation on the energy efficiency of hydroelectric stations.

Zaidi and Khan (Zaidi and Khan, 2018) addressed the importance of exploring alternative energy sources, such as run-of-the-river hydropower projects. They proposed a novel approach utilizing geospatial data and DEMs to enhance the siting process.

Huang and Yan (Huang and Yan, 2009) provide a comprehensive description of the hydropower station distributions in China, indicating the dependence of elevation and hydropower potential and development.

Bartle (Bartle, 2002) advocated the global potential of hydropower, showing the possible interplay between elevation and possible hydropower development.

Rojanamon (Rojanamon et al., 2009) emphasized the use of GIS in siting small run-of-the-river hydropower stations, in which elevation serves as a critical role, with other factors, such as engineering, economic and environmental factors.

Yizhi Tian (Tian et al., 2020) studied the usage of Geographical Information Systems to assess hydroelectricity generation plans in terms of potential and cost.

### 3.2 Impact of Precipitation on Hydroelectric Stations

Multiple studies and articles point towards the incredible importance of precipitation levels in hydroelectric power generation, and as a direct result the importance of the location in which the power station is built.

Wei, Li, et al. (Wei Li, 2020) identify the large effect precipitation has on hydroelectric power generation, and quantified the effect seasonal variation in precipitation has on current hydroelectric power generation. Additionally, they modeled the expected variation in rain for the future and in turn the effect on hydroelectric power generation providing decision support for future water resource management.

Senni and von Jadow (Senni and von Jagow, 2023) explain the large effects of precipitation on hydroelectirc power generation, specifically the association between more variable or risky precipitation locations and lower power generations.

### 3.3 Impact of Power Consumption Locality on Hydroelectric Stations

Many works highlight the importance of power generation in areas where it is most convenient and necessary for those nearby to meet living standards.

Azad et al. (Azad et al., 2020) focuses on the impact that renewable energy sources, primarily hydropower, have on developing countries with expanding standards of living. They determine that hydropower has reached high levels of technical sophistication.

Hoes et al. (Hoes et al., 2017) study how the increasing energy demand due to population growth creates urgent demand for more sources of hydropower. Their research provides a detailed evaluation of hydropower potential in several locations.

## 4 Datasets

The table 1 illustrates the schema of our datasets. We joined them by using combined key of longitude and latitude.

| Datasets | Attributes | |
|---|---|---|
| Terrain Tiles | longitude | String |
| | latitude | String |
| | elevation | Double |
| NCEI GSOY | Station | String |
| | Longitude | Double |
| | Latitude | Double |
| | Elevation | Double |
| | Precipitation | Double |
| EIA RECS | State | String |
| | Classification | Char |
| | Energy Usage | Double |
| Cities | City | String |
| | Longitude | Double |
| | Latitude | Double |
| | Population | Integer |

Table 1: Datasets schema

### 4.1 Terrain Tiles

The first dataset of our paper is an AWS S3 open data registry called Terrain Tiles. The dataset are consist of tif terrain tiles in 15 different zoom levels. The higher the zoom levels, the more locations, in terms of longitude and latitude. Each gridded

elevation tiles in the tif files contains the mean elevation. In our analysis, we chose zoom level 4 for convenience. The world map is divided by the grid with total 16 squares (4*4). Then, each square is further divided into 16 sections, and each section is stored in a tif file, which essentially is a picture of one of many components of the world map. To get the actual dataset, we used special python packages osgeo to extract the longitude, latitude and elevation data and integrate them into one csv file.

### 4.2 NCEI Global Summary of the Year

Our second dataset consists of weather data gathered globally by the National Center for Environmental Information. This dataset records precipitation and wind data yearly from weather stations across the globe. Records stretch back as far as the 1700s and the dataset is continually updated to contain the most recent information. While each weather station did not necessarily record the same information, weather stations recorded data under the same column headers. For our purposes we were mostly interested in data under the column header "PRCP" which corresponds to the total precipitation recorded along with Station ID, Latitude, Longitude, and Elevation.

### 4.3 EIA REC Survey

The number of relevant datasets available for analyzing power consumption is scarce. We settled on using the microdata from the 2020 US Energy Information Administration RECS as a third dataset. This data consisted of individual households and their energy consumption. It also contained countless pieces of extraneous data on power consumption from individual appliances, which we had to clean out.

### 4.4 Cities Data

We also used a final dataset of US cities, their coordinates, and their populations from SimpleMaps in order to consolidate our data.

## 5 Analysis

### 5.1 Elevation Data Analysis

For this dataset, we first extract the data by using python library, which is mentioned in 3.1. Then, we deleted rows that contains null values or outliers in all three columns. In addition, we checked the validity of these 3 columns since they have specific range, and excluded only the coordinates in the United States by cutting the longitude and latitude. To align with other dataset for further analysis, we modified the longitude and latitude by 2 decimal places. Note that although by taking 2 decimal places, we calculated the elevation by taking the mean of any location with the 2 decimal places accuracy, there are still over 900k rows for this dataset. Therefore, we can still guarantee the goodness of the dataset.

### 5.2 Precipitation Data Analysis

When first downloaded, the NCEI Global Summary of the Year data set consisted of approximately 100,000 CSV files each representing a weather station. As our analysis is focused on the United States, the weather stations outside of the USA were filtered out leaving 34,000 remaining files. Then using the file system, the remaining CSV files were concatenated into one singular file and uploaded to HDFS. Cleaning of the weather data largely consisted of reformatting the many concatenated files to be consistent and discarding extraneous data not required for our analysis. In the concatenated CSV, each header line represented a weather station with all following lines until the next header representing a year of data recorded at that weather station. Processing consisted of averaging the precipitation recorded over the years for a given weather station and combining it into one line. The end result was a schema as shown in Table 1; a CSV file with 5 columns, Station, Longitude, Latitude, Elevation, and Precipitation. Profiling the cleaned data was relatively simple with mean, median, mode, and standard deviation of the average precipitation being calculated across all weather stations in the USA.

### 5.3 Energy Consumption Analysis

The RECS data consisted of around 18,000 individual household statistics. These statistics included energy usage in BTU/hr, the state, the urban/rural classification, and extraneous data on appliance power consumption. There is no data on the specific locations of these households. To make use of this data, we conducted a few map reduce jobs to calculate the mean power consumption for rural, cluster, and urban areas. The data concluded that there is no significant difference in energy consumption per household in the three different classifications of cities. We then consolidated the population data of cities with the topographical and

precipitation data.

### 5.4 Data Merging and Scoring

In order to perform analytics we needed to merge our three data sources into one table which was done with by joining on the latitude and longitude in each data source; the result was a table with each row identifying a possible location with latitude and longitude. Each location had the attributes of precipitation, elevation, and population. With the data finalized, we proceeded to use a scoring function to assign a score to each location; the higher the score, the better we found a hydroelectric station to perform in that location. The scoring function has a weight of 2.0 assigned to precipitation because it is the most important factor in how much energy a hydroelectric plant can produce as shown by [Wei, Li, et al. (2020); Senni and von Jagow (2023)]. Similarly we assigned a weight of 1.0 to elevation because, while still an important factor in how much energy is generated, it does not have as great an impact as precipitation according to [Huang and Yan (2000); Bartle (2002)]. Finally, population was normalized by diving by 10,000 and then a weight of 0.3 was applied as we found long distance power transfer loss is quite low, so the population of the area is not as important of a factor in how much power can be utilized. With the final score calculated, we normalized the score by computing a Softmax so that all scores were between 0 and 1, with 1 being the best location to place a hydroelectric power station and 0 being the worst.

$$\mathbf{S = 2.0 \times Precipitation + 1.0 \times Elevation + 0.3 \times Population}$$

## 6 Visualizations

After deriving the normalized score for each individual locations, we further examined the correlation between the scores and potential factors.

### 6.1 Power station heat map

Figure 2 is the heat map of normalized scores for each location in the united states. Points on the map are color-coded to represent the scores, from purple (low) to yellow (high). The geographical distribution suggests a higher concentration of preferable locations in the eastern half of the country, while the west shows a more sparse and varied score distribution. This might indicate the regional differences, such as climate, infrastructure, and population density.

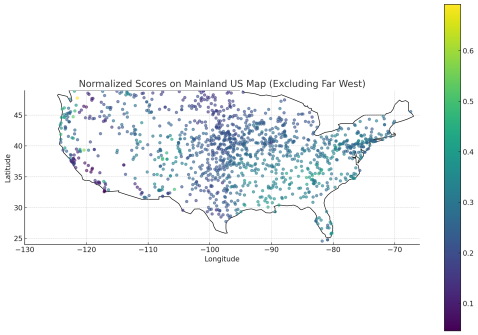


Figure 2: Score heat maps in USA

### 6.2 Factor Analysis

**Score vs. Elevation** As we can see from Figure 3 that there is a position correlation between the elevation and normalized scores. This indicates that locations at higher elevations are more suitable for building power stations. This tread implies the potential elevation related factors, such as climate or accessibility, could influence the power output. As for the limitations, it is important to note that the distribution of data is imbalanced, and there are some outliers at extreme elevations, which could be affected other unexplored factors.

**Score vs. Average Precipitation** The correlation between precipitation and our scores seems to be more relevant. Although it seems to be a subjective common sense, our analysis further confirms the relationships. The pattern are clear, with some variations at the low precipitation areas, which could be affected by hidden variables.

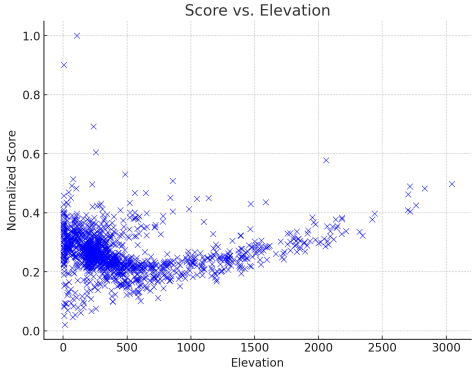


Figure 3: Score vs. Elevation

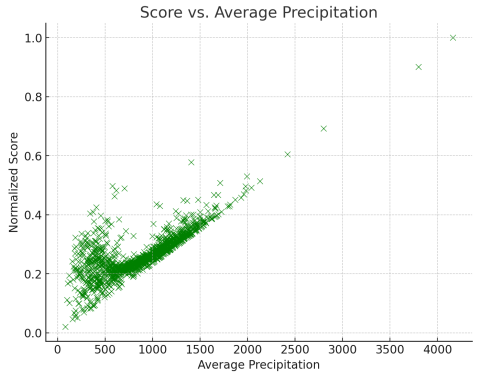


Figure 4: Score vs. Precipitation

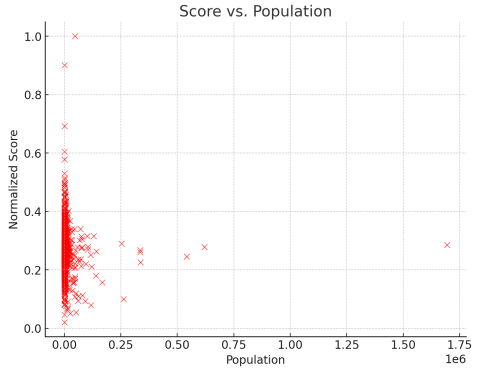


Figure 5: Score vs. Population

**Score vs. Populations** we assume the distribution of energy usage could be represented by population. However, the result in the graph does not reveal a pattern between the population and the score. This could because of the false assumption between the cities and population. Also, the result might be improved if more energy distribution factors can be considered.

## 7 Conclusion

In summary, in this paper we proposed a solution to approximate the optimal locations for building hydroelectric stations in terms of longitude and latitude. Furthermore, Our analysis utilized 3 datasets from different fields and joined them together to find potential influential factors for power station locations. After merging the dataset together, we also calculated the normalized scores as well as generating visualizations. We would like to present our analysis as a guide to help determining optimal hydroelectric power station locations, as well as inspiring future work to explore more potential factors that might have impacts on saving energy. Also, since we only measured certain locations in the United States, the relationship between the normalized scores and the factors might not be the same in other regions in the world.

## Acknowledgement

Many thanks to Professor Ann Malavet and NYU HPC for their help.

## References

Mohammad Askari, Vahid Mirzaei Mahmoud Abadi, Mohsen Mirhabibi, and Parvin Dehghani. 2015. Hydroelectric energy advantages and disadvantages. *American Journal of Energy Science*, 2:17–20.

Abdus Samad Azad, Md Shokor A. Rahaman, Junzo Watada, Pandian Vasant, and Jose Antonio Gamez Vintaned. 2020. Optimization of the hydropower energy generation using meta-heuristic approaches: A review. *Energy Reports*, 6:2230–2248.

Alison Bartle. 2002. Hydropower potential and development activities. *Energy Policy*, 30(14):1231–1239.

Olivier A. C. Hoes, Lourens J. J. Meijer, Ruud J. van der Ent, and Nick C. van de Giesen. 2017. Systematic high-resolution assessment of global hydropower potential. *PLOS ONE*, 12(2):1–10.

Hailun Huang and Zheng Yan. 2009. Present situation and future prospect of hydropower in china. *Renewable and Sustainable Energy Reviews*, 13(6-7):1652–1656.

Pannathat Rojanamon, Taweep Chaisomphob, and Thawilwadee Bureekul. 2009. Application of geographical information system to site selection of small run-of-the-river hydropower project by considering engineering/economic/environmental criteria and social impact. *Renewable and Sustainable Energy Reviews*, 13(9):2336–2348.

Chiara Colesanti Senni and Adrian von Jagow. 2023. Water risks for hydroelectricity generation. *Centre for Climate Change Economics and Policy Working Paper 418/Grantham Research Institute on Climate Change and the Environment Working Paper 394. London: London School of Economics and Political Science*.

Yizhi Tian, Feng Zhang, Zhi Yuan, Zihang Che, and Nicholas Zafetti. 2020. Assessment power generation potential of small hydropower plants using gis software. *Energy Reports*, 6:1393–1404.

Junhong Guo Zhe Bao Lingbo Fu Baodeng Hou Wei Li, Jiheng Li. 2020. The effect of precipitation on hydropower generation capacity: A perspective of climate change. *Frontiers in Earth Science*, 8.

A. Z. Zaidi and M. Khan. 2018. Identifying high potential locations for run-of-the-river hydroelectric power plants using gis and digital elevation models. *ideas.repec.org*.