Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements

Walter P Moore

Henry Hughes

4A Civil Engineering

May 2023

534 Crimson Court

Oshawa ON L1J 8E2

May 16, 2023

Dr. Rania Al-Hammoud

Associate Chair, Undergraduate Studies

Department of Civil and Environmental Engineering

University of Waterloo

Waterloo, ON, N2L 3G1

Dear Dr. Al-Hammoud:

This report, entitled "Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements" was prepared as my 3A work term report. During my work term, I worked on projects dam reinforcement projects. These dams needed to be reinforced due to higher levels of discharge that initially expected. The purpose of this report is to gain additional understanding into how levels of hydrological discharge can be forecasted by using data driven methods.

This report was prepared during my Co-op at Walter P Moore. I would like to thank the team at Walter P Moore, as they allowed me to work on several dam projects which inspired me to write a report that is focused on how data driven methods for hydrological forecasting can be used to create more resilient infrastructure.

This report focuses on the numerical methods that are multiple linear regression (MLR), and extreme learning machines (ELM). While not being the most complicated mathematical tools that can be utilized for this type of analysis, these methods are not too challenging to apply, and allow for standard datasets to be used without much difficulty.

This report was written entirely by me and has not received any previous academic credit at this or any other academic institution. I received no outside help with the report.

Sincerely,

Text, letter

Description automatically generated

Henry Hughes  
ID# 20792437

Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements

Walter P Moore

Henry Hughes

4A Civil Engineering

May 2023

# Summary

The aim of the Summary is to convey the main points of a report to senior management personnel who are not usually involved in technical details. For this reason, it is sometimes labeled Executive Summary. It should be written last and must be written in full sentence form while avoiding too much jargon or heavy details. The Summary may be thought of as a mini report of about 250 words (3-4 paragraphs) that must be able to be read and understood in isolation from the main body of the report. It must not exceed one page and should include the scope and purpose of the report (including background/rationale); the technical approach taken, which might include a discussion of the alternatives considered and the major technical findings; and the principal conclusions and recommendations. It must not include tables and figures or a direct reference to tables, figures, and sections contained in the main report.

# Acknowledgements

I wish to thank Professor John Quilty for allowing me to be a part of his research team, which has inspired me to write about this topic. The work his team does is fascinating, and the reason I am focused on learning more about the topic of applied machine learning and statistical analysis in hydrology.

**Table of Contents**

[Summary v](#_Toc132575074)

[Acknowledgements vi](#_Toc132575075)

[List of Figures viii](#_Toc132575076)

[List of Tables viii](#_Toc132575077)

[List of Appendices viii](#_Toc132575078)

[1.0 Introduction 1](#_Toc132575079)

[1.1. General 1](#_Toc132575080)

[1.2. Background 1](#_Toc132575081)

[1.3. Scope and Objectives 1](#_Toc132575082)

[2.0 Multiple Linear Regression 2](#_Toc132575083)

[2.1. Multiple Linear Regression Formulas 2](#_Toc132575084)

[2.2. Variable Dependency 3](#_Toc132575085)

[3.0 Extreme Learning Machine 3](#_Toc132575086)

[4.0 Development of Dataset 3](#_Toc132575087)

[4.1. Dataset of Choice 4](#_Toc132575088)

[4.2. Data Cleaning 5](#_Toc132575089)

[5.0 Application of and Analysis of MLR 7](#_Toc132575090)

[5.1. Data Splitting 8](#_Toc132575091)

[5.2. MLR Results 9](#_Toc132575092)

[6.0 Recommendation (optional) 10](#_Toc132575093)

[References 12](#_Toc132575094)

# List of Figures

[**Figure 1:** Sample summary page 6](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472653)

[**Figure 2:** Sample of a Table of Contents with more than three appendices and lengthy lists of tables and figures 7](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472654)

[**Figure 3**: Example of a table included in a work report. It should be noted that the table is presented immediately after it is first introduced in the text. 8](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472655)

[**Figure 4:**Example of an equation layout in a work report 9](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472656)

[**Figure 5:** Example of a citation for a table taken from a published source 14](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472657)

[**Figure 6:** Example of a citation for a figure and data taken from a published source 14](https://uofwaterloo-my.sharepoint.com/personal/dlacroix_uwaterloo_ca/Documents/Work%20Reports%20-%20CEE%20UW/CEE%20WKRPT/Work%20Report%20Guidelines/Technical%20Report/Template%20for%20technical%20report%20draft%202.docx#_Toc69472658)

# List of Tables

[**Table 1:** Sample Table 9](#_Toc69472344)

# List of Appendices

[**Appendix A**: Explanation of Symbols Used for Marking Work Report Grammar. 18](#_Toc69472458)

[**Appendix B:** Avoiding Common Grammar Errors 19](#_Toc69472459)

[**Appendix C:** Proofreading Checklist 20](#_Toc69472460)

[**Appendix D:** Writing Tips 21](#_Toc69472461)

# Introduction

## General

There are many benefits to understanding how water moves. It is also beneficial to understand how the movement of that water changes over time. It allows cities to be constructed in the right locations, and with the correct water resources asset, ensures proper flood protection measures can be put in place to protect critical infrastructure, and will allow society to function in a way that allows people to utilize the benefits of water, without the drawbacks. This is a challenge that engineers must face, and use technology to better understand, so that moving forward, infrastructure can be created in the most effective way possible to serve individuals around the world.

## Background

Water resources assets are critical for any society to function. This primarily includes the rivers, lakes, and groundwater from which we obtain all our potable water. It also includes physical assets such as dams, water and wastewater treatment plants, desalination plants, and water distribution networks. To ensure these physical assets will be able to last into the future, their owners must be able to forecast their future demand. For dams, the demand is how much upstream discharge will pass through the dam. Accurately forecasting this discharge will allow dams of the correct size to be constructed, increasing their longevity.

## Scope and Objectives

There are several ways that discharge through a dam can be forecasted. Historically, physical methods have been used. These were primarily a function of historical precipitation, and the geography of a specific river basin. This would allow engineers to forecast discharge through dams so their size could be determined. Recently, data driven methods have been used to forecast hydrological discharge. These methods are more efficient, since they can be done with basic data, and physical methods that require lots of empirical data are difficult to come up with, particularly in a short amount of time. This report will detail and apply two data driven methods; 1. Multiple Linear Regression, and 2. Extreme Learning Machines. The results will then be analyzed, and future methods will be explored.

# Multiple Linear Regression

Multiple linear regression is a simple statistical tool that is used to create forecasts for a specific variable of interest, based on a variable number of explanatory variables. In essence, it identifies the linear relationship between a set of input explanatory variables and an output variable. The mathematical model allows the output variable to be forecasted into the future by utilizing a linear equation that projects the dependent variable.

## Multiple Linear Regression Formulas

The general equation for MLR is as follows:

In the above equation, y is the dependent variable, or the output. This variable can be linked to the n independent variables, which are represented by xi. Each of these independent variables are weighted by their coefficients ai. The coefficient ao is a correction term, which allows the curve to be properly fitted. All the other coefficients are what define the relationship between each independent variable xi and the dependent variable y.

To determine whether an independent variable is correlated to the dependent variable in a significant way, a p-value can be used. In linear regression, a p-value of one means that the independent variable in question is perfectly correlated with the dependent variable. A p-value of zero means there is no correlation between the independent and dependent variables. The p-value is calculated based on a t-test of the estimated coefficient, which compares the estimated coefficient to the standard error of the estimate.

The goal of MLR is to determine the values for the coefficients ai such that the function in equation one above best fits the data, which is typically done by minimizing the sum of the squared errors between the actual values of the dependent variable and the predicted values of the dependent variable. This can all be done using basic tools such as Microsoft Excel or Python.

## Variable Dependency

A benefit to MLR is that there is not any connection between the independent variables in the equation. However, it is important to analyze the variables in a way that results in a model that is statistically efficient. This means that multicollinear variables should not be included in the model, and all variables should be statistically significant.

## Common MLR Use Cases

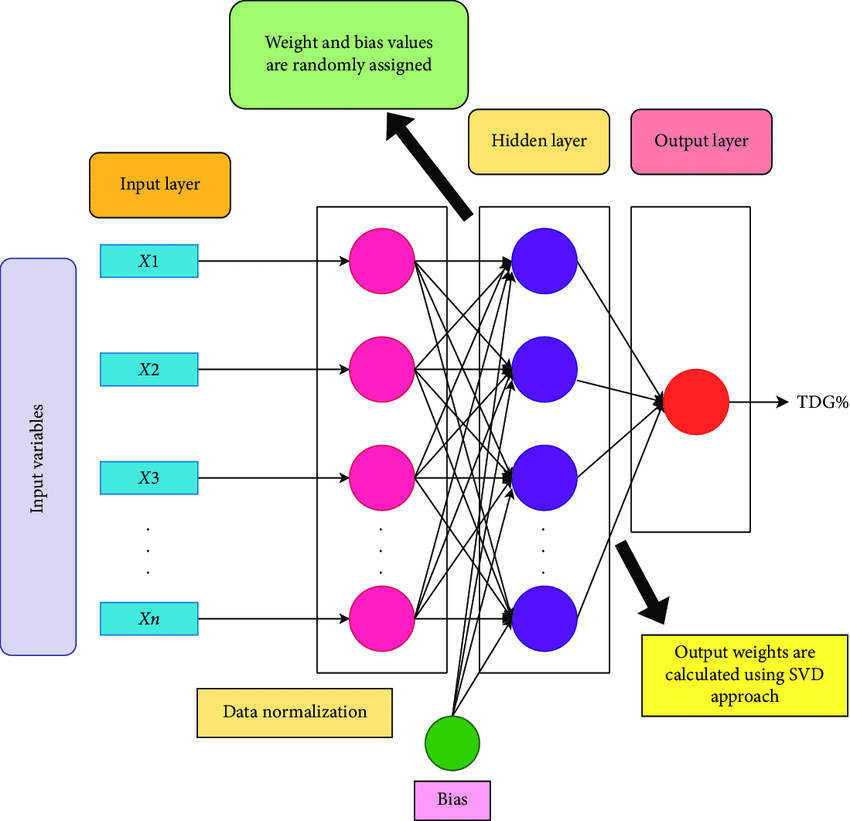
MLR is a common statistical technique that has been used for a wide variety of applications. One reason for this is that MLR is a very accessible tool that almost anyone can apply if they have a complete dataset. Microsoft Excel has a built-in set of tools for using MLR, meaning that it is often a tool people turn to when they have a dataset and are looking to create a meaningful forecast with it.

# Extreme Learning Machine

An ELM is an algorithm that can be used to develop data forecasts. It is a single layer feedforward neural network. The idea behind an ELM is to randomly select a set of hidden neurons and computer their output weights analytically. ELM algorithms are usually used for regression-based models. While being quite straightforward, ELM has been shown to be an effective technique for forecasting. The main benefit of the ELM method is that it is computationally efficient since it does not require backpropagation to train the neural network.

## ELM Mathematics

An ELM has a few different components, including an input layer, a layer of hidden neurons, and an output layer. The general structure can be seen in the Figure below (https://www.researchgate.net/figure/Extreme-learning-machine-ELM-structure\_fig1\_348369700):



The steps to determine the weights can be summarized as shown in the following list: 1. Multiply inputs by weights, 2. Apply the bias, 3. Apply an activation function, 4. Calculate the output, and 5. Apply the matrix inverse method (https://towardsdatascience.com/introduction-to-extreme-learning-machines-c020020ff82b). After this, the output can be computed easily by using the following equation:

The actual application of the ELM algorithm is straightforward when using Python, as there are packages that exist which will do the computing of the ELM algorithms.

## ELM History and Prior Applications

ELM was first proposed by Huang et al in 2006. It was proposed since it is a much more computationally efficient method than other neural network algorithms. ELM has been used in a wide variety of applications, including image classification, and sales predictions, with the main benefit being its fast-computing time. ELM has been used frequently in hydrology, most notably by Professor John Quilty at the University of Waterloo. Several papers have been published by him that focus on ELM and other machine learning techniques for streamflow forecasting. Including <https://www.sciencedirect.com/science/article/pii/S0022169416305893#f0010>.

# Development of Dataset

To apply the MLR and ELM methods described in the prior sections of this report, one needs data that can be used to create a forecast. I was unable to use data from the project I worked on at Walter P Moore. That specific project focused on data for the Brazos River in Texas and included things such as rainfall and discharge amounts from areas near the dam of interest.

## Dataset of Choice

To test the MLR and ELM methods, a dataset was used that was found online. The data was like what was used for the project that Walter P Moore completed. To analyze the data, Python was used, along with the Pandas and Matplotlib packages.

The data is taken from Spain. There are three rainfall measurements, and six discharge measurements. The Gramenet discharge lies below the rest of the measurements of interest in terms of elevation, therefore it is the variable of interest. This is because the water from the other discharges and the precipitation will contribute to the Gramenet discharge. The other variables will be used to forecast it.

Table 1 shows the data from the first four explanatory variables, and Table 2 shows the data from the final five explanatory variables. They have been split so the tables remain readable. The data continues until December 31, 2010, and it can be seen in Appendix A.

| **Date** | **Barcelona\_Fabra Daily Rainfall [mm]** | **Barcelona Daily Rainfall [mm]** | **Sabadell\_Aero Daily Rainfall [mm]** | **Garriga Discharge [m^3]** |
| --- | --- | --- | --- | --- |
| 1/1/2003 | 0 | 0 | 0 | 0.254 |
| 1/2/2003 | 0 | 0 |  | 0.254 |
| 1/3/2003 | 0 | 0 |  | 0.246 |
| 1/4/2003 | 0 | 0 | 0 | 0.251 |
| 1/5/2003 | 18.1 | 16.4 | 2.1 | 0.241 |
| 1/6/2003 | 1 | 0.2 | 4.3 | 0.644 |
| 1/7/2003 | 0 | 0.5 |  | 0.317 |
| 1/8/2003 | 2.1 | 1.1 |  | 0.314 |
| 1/9/2003 | 0 | 0 |  | 0.309 |

| **Date** | **Castellar Discharge [m^3]** | **Llica Discharge [m^3]** | **Montornes Discharge [m^3]** | **Mogoda Discharge [m^3]** | **Gramenet Discharge [m^3]** |
| --- | --- | --- | --- | --- | --- |
| 1/1/2003 | 0.0327 | 0.155 | 0.3985 | 0.12 | 3.6275 |
| 1/2/2003 | 0.0281 | 0.151 | 0.3745 | 0.106 | 3.9471 |
| 1/3/2003 | 0.0225 | 0.145 | 0.375 | 0.101 | 4.0174 |
| 1/4/2003 | 0.03 | 0.145 | 0.372 | 0.099 | 3.9316 |
| 1/5/2003 | 0.0328 | 0.146 | 0.3686 | 0.097 | 3.7913 |
| 1/6/2003 | 0.0559 | 0.205 | 1.0558 | 0.516 | 7.6332 |
| 1/7/2003 | 0.039 | 0.168 | 0.6 | 0.189 | 4.6152 |
| 1/8/2003 | 0.0386 | 0.163 | 0.519 | 0.159 | 4.7008 |
| 1/9/2003 | 0.0315 | 0.158 | 0.4775 | 0.138 | 4.4481 |

A sample graph of the Barcelona\_Fabra Daily Rainfall [mm] time series can be seen in the Figure below. The rest of the graphs for the input variables can be seen in the Appendix.

Chart

Description automatically generated

## Data Cleaning

As can be seen in Tables 1 and 2, the data is not perfect. There are many missing values. It is important to understand what missing values mean, so they can be dealt with in a reasonable manner. These are easy to evaluate in Python. A heatmap can be used to show how many values are missing. This can be seen in the Figure below:

A picture containing chart

Description automatically generated

As can be see in the Figure above, the Castellar and Montornes measurements are missing a lot of data. As a result of this, they have been entirely removed from the dataset. This leaves the dataset with two less explanatory or input variables. The new data structure can be seen in the Table below.

| **Date** | **Barcelona\_Fabra Daily Rainfall [mm]** | **Barcelona Daily Rainfall [mm]** | **Sabadell\_Aero Daily Rainfall [mm]** | **Garriga Discharge [m^3]** | **Llica Discharge [m^3]** | **Mogoda Discharge [m^3]** | **Gramenet Discharge [m^3]** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1/1/2003 | 0.0 | 0.0 | 0.0 | 0.254 | 0.155 | 0.120 | 3.6275 |
| 1/2/2003 | 0.0 | 0.0 | NaN | 0.254 | 0.151 | 0.106 | 3.9471 |
| 1/3/2003 | 0.0 | 0.0 | NaN | 0.246 | 0.145 | 0.101 | 4.0174 |
| 1/4/2003 | 0.0 | 0.0 | 0.0 | 0.251 | 0.145 | 0.099 | 3.9316 |
| 1/5/2003 | 18.1 | 16.4 | 2.1 | 0.241 | 0.146 | 0.097 | 3.7913 |

Finally, any rows that included null data were removed. This resulted in 2076 data points that can be utilized to run the simulations. To properly run MLR and ELM simulations, complete datasets are key. Even though the dataset used was not complete initially, modifications were done to make the dataset complete and ready for evaluation. The code that was written to clean the dataset can be seen in Appendix A.

# Application of and Analysis of MLR

MLR is not a technique that can be used on all datasets. For MLR to be effective, there needs to be a linear or consistent relationship between the explanatory or input variables, and the output variables. The reason for this is quite simple, and can be demonstrated by analyzing the equation below:

For example, if the output variable y is streamflow discharge, and the input variable x1 is the rainfall in a nearby basin, then as the rainfall increases, the discharge will also increase, assuming the coefficient for the input variable x1 is positive. Therefore, for MLR to be effective, the input variables need to affect the output variable in the same way throughout the dataset. This can be checked by plotting the input variables against the output variable. The Figure below shows the relationship between the input variable Barcelona\_Fabra and Gramenet Discharge variables. The Gramenet variable is in orange.

Chart

Description automatically generated

Generally, as the input variable goes up, so does the output variable. This means that MLR will be applicable to this input variable. Overall, all six of the input variables have similar relationships to the output variable. This is shown in the Appendix. Therefore, MLR can be applied to all the variables in question.

## Data Splitting

To apply MLR to a dataset, it is important to split the data. The reasoning for this is quite simple. An equation will be produced that allows the output variable to be forecasted using the input variable. It is important to think about how the equation can be tested. The best way is to have additional data where the input variables can create a forecast, and the forecast can be compared against the actual results. This can be done by splitting the existing data into two groups, which include 1. Training data, and 2. Test data. The training data can be used to generate the coefficients, and the test data can be used to check how effective the MLR equation was.

In Python and Pandas, it is easy to split the dataset. The following code was used, where x and y are the input and output variables that are being split, respectively.

x\_train, x\_test, y\_train, y\_test = train\_test\_split(

    x, y, test\_size=0.2, random\_state=101)

It should be noted that random state means that the results will be reproducible if the code is run again. The test size is the amount of data that is used for testing, so 20% of the data is used for testing and 80% is used for training.

## MLR Results

In Python, there are packages that allow MLR simulations to be easily completed. The following code was used to generate predictions:

# Create a linear regression model and train it using the training data.

model = LinearRegression()

model.fit(x\_train, y\_train)

predictions = model.predict(x\_test)

The model was trained based on the input matrix x\_train which contained independent variables, and a vector y\_train which contained dependent variables. The model was used to forecast the output variables for the set of data x\_test. To see how accurate the model was, it can be plotted as shown in the Figure below.

Chart, line chart, histogram

Description automatically generated

As the figure shows, the forecast is fairly accurate. This can be further quantified with simple metrics such as… insert here.

# Application and Analysis of ELM

# References

Include Kaggle data, some of Quilty’s papers.

**Appendix A:** Explanation of Symbols Used for Marking Work Report Grammar.

Please see separate document on Learn/Piazza.

Appendix B: Avoiding Common Grammar Errors

Please see separate document on Learn/Piazza.

Appendix C: Proofreading Checklist

Please see separate document on Learn/Piazza.

Appendix D: Writing Tips

Please see separate document on Learn/Piazza.