Kristianstad University
SE-291 88 Kristianstad
+46 44 250 30 00
www.hkr.se

Independent project (degree project), 15 credits, for the degree of Bachelor of Science (180 credits) with a major in Computer Science
Spring Semester 2021
Faculty of Natural Sciences

Effective machine learning techniques for detecting Inflow and Infiltration water in wastewater channels

Jonas Johansson and
Viktor Westlund

Author

Jonas Johansson and Viktor Westlund

Title

Effective machine learning techniques for detecting Inflow and Infiltration water in wastewater channels

Supervisor

Qinghua Wang and Magnus Lindgren

Examiner

Kamilla Klonowska

Abstract

In this degree thesis, the research team analyses different ways of monitoring water flow in wastewater channels finds an effective way of processing data using machine learning on collected data from a specific area and sets up a system for finding correlations between rain levels, water usages and wastewater flow to detect infiltrations and inflow water in wastewater channels. The study concludes that a sensor network is most suited for monitoring wastewater channels and that machine learning could be used to detect infiltration and inflow water using different types of data.

Keywords

Infiltration and inflow, machine learning, sensor network, wastewater system

Preface

The exam work has been a 15 ECTS-points project conducted for Kristianstad kommun water- and sewer system department.

We as a research team would like to thank our supervisors Qinghua Wang at Högskolan Kristianstad and Magnus Lindgren at Kristianstad kommun for their amazing support throughout the project, for continuously helping us with everything from machine learning to ways to approach this project, and for giving us the knowledge and understanding needed to conduct this research. We would also like to thank Hans-Inge Hansson (Kristianstad kommun), Johan Holmberg (Atea) and Jonas Månsson (Kristianstad kommun) among others, for all the help and interest in our work.

Contents

Keywords and Abbreviations

Additional water (could also be mentioned as I/I-water) - water that in some way enters the wastewater system that does not belong there (rainwater for example)

Deep learning - "Deep Learning is a machine learning technique that constructs artificial neural networks to mimic the structure and function of the human brain. In practice, deep learning, also known as deep structured learning or hierarchical learning, uses a large number hidden layers - typically more than 6 but often much higher - of nonlinear processing to extract features from data and transform the data into different levels of abstraction (representations)." [8]

I/I (Infiltration and Inflow) water - a collection of ways the additional water enters the wastewater channel. Infiltration is water that enters the pipe through cracks or otherwise broken pipes, while inflow mainly means wrongly connected pipes.

MDPI - Multidisciplinary Digital Publishing Institute is a publisher of academic papers through journals

Stormwater/drainage system - channels (system of pipes) that deal with water from rain, snow, and meltwater from ice.

SWMM - Storm Water Management Model is an application that is used to simulate and analyze water management systems

Wastewater system - wastewater channels that go to the sewage treatment plant

1. Introduction

1.1 Background

What is considered to be the world's first Internet of Things (IoT) device was introduced in 1990 when John Romkey had created a toaster which could be operated from a computer using TCP/IP networking and simple network management protocol (SNMP) to access the management information base (MIB) [1]. In 1999 "Internet of Things" term was brought to life for what is seen as the first time, by Kevin Ashton who was working in supply chain optimization in Procter and Gamble [2]. By 2003, 500 million devices were connected to the internet, and further on what scientists mean is that the true birth of the Internet of Things era is somewhere between 2008 to 2009 with the definition; "the time where more devices were connected to the internet than humans alive" [3].

Though it is difficult to specify the exact number, the estimated number of connected IoT devices by 2021 is somewhere between 20 to 30 billion worldwide [4].

Based on this, it cannot be seen as a coincidence that it is more and more common in today's society to use IoT to track and contain data in all kinds of areas. A report made by the National League of Cities shows that over 66% of the cities in the United States were investing in smart-city solutions in 2017 [5]. Some of the most common smart-city priorities were public wifi-areas, intelligent traffic signals, e-governance applications, and smart-meter utilities. The IoT implementations make it easier for governments to gather data and create statistics from different sources. The gathered data is then often used to analyze different occurrences and how certain impacts, scenarios, and anomalies can affect the entirety. Though the heavy growth for the use of IoT devices not only in our homes and our everyday-use items but in society all around us, an area where the implementation of IoT devices is mostly still under development, prototyping, and implementation is within the sewage management sector [14] [3]. Using sensors

in drainage systems can have many benefits including real-time monitoring of water flow to quickly locate and find any potential leakages in the system and benefits the safety of workers that otherwise need to physically enter sewages for manual measurements. Sensors also provide more frequent data gathering that makes it easier to continuously analyze the incoming data. This could also benefit the optimization of the whole drainage system by being able to analyze statistics to find anomalies and correlations in different occurrences. For example, being able to find places where water levels in wastewater channels go up on days where it rains could mean that there are leakages from the stormwater pipe into the spill water pipe. A study conducted in Poland focused on stormwater entering the wastewater system by analyzing the water entering the sewage treatment plant of Nowy Targ as well as measuring the rainwater. The research showed that the amount of stormwater in the sewage system was 14.2% of the total sewage water on days where the rain was less than 5 millimeters, and 32.5% of the total sewage water on days where it rained above 20 millimeters [6]. The municipality of Stockholm decided in 2016 to rebuild the sewage treatment plant "Henriksdalsreningsverk" to double the processing power to accommodate more households to be connected. The rebuild of the treatment plant cost roughly $700 million (6 billion SEK). A report released by Stockholm Vatten och Avfall in 2020, showed that the amount of inflow and infiltration water in the municipality of Stockholm is measured to be 40% of the total amount of water that is handled by the treatment plants [7].

1.2 The Inflow and Infiltration Problem

Unwanted water entering the sewage pipes can be defined as inflow and infiltration water. Inflow water is water that enters the sewage pipes through misconnected junctions, for example where rainwater should instead be led to drainage pipes. Infiltration water is instead water that through cracks enters the sewage pipes. An example of this can be found in the attached figure 20 below.

Figure 20. Infiltration and Inflow [25].

This can be groundwater pressing on the pipes from underneath or water from leaking drainage pipes as well as leaking drinkwater pipes. In figure 19 an example is shown of how the drainage and sewage pipes are located in terms of levels.

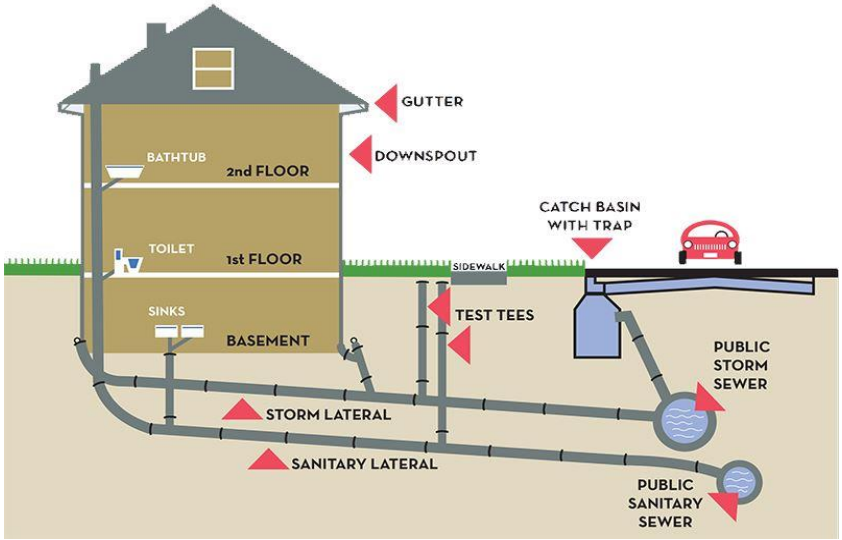

Figure 19. Structure of drainage and sewer pipes [24].

1.3 The Focus of the Study

The purpose of the experiment was to come up with an effective way on how to analyze and find infiltration and inflow-water in data gathered from water-level-, flow-, rainfall- and water supply sensors by using machine learning techniques to process the already collected data from a specific area. The goal of the experiment is then to simulate the predictions in SWMM.

The purpose of the literature review is to find the existing solutions for monitoring and handling data analysis in sewage systems by comparing the most common ones.

1.4 Problem Statement

By finding an effective and efficient way of monitoring sewage systems, leakages, infiltrations and inflow, as well as misconnected pipes, can easily be localized and dealt with. This will help optimize the sewer system's capacity and lower the amount of unnecessary I/I water the sewage treatment plants process.

1.5 Thesis Structure

1.5.1 Methodology Introduction

The thesis has been conducted using a literature review and by conducting a case study. In the literature review part, the research team has searched for relevant literature to answer the research question; "How do monitoring and data analysis in sewage systems compare and which techniques are the most effective?". This result the team later used to base the case study on. The case study was conducted using different machine learning algorithms to analyze sets of data.

1.5.2 Overall Structure

Different techniques have so far been implemented in water and sewer systems. In this thesis, the authors of this thesis focused on finding infiltration and inflow to wastewater channels. In the literature review, the most common ways of

implementing a monitoring system for sewage pipes as well as different ways of handling and interpreting the collected data are presented. The results from the literature review are discussed and evaluated along with a conclusion of the data found. In the experiment's result part of the thesis, the research team has presented the outcome of the data handling that has been done on data collected from different sensors. The processed data is then examined in the discussion part of the thesis and is concluded in the conclusion.

2 Methodology

2.1 Literature Review Methodology

2.1.1 Data Collection

As many research papers were found that included parts of the topic of research, the research team only chose to include the most distinctive resources compared to the chosen topic.

2.1.2 Literature Review Research Question

How do monitoring and data analysis in sewage systems compare and which techniques are the most effective?

2.1.3 Search Phrases

Optimize “wastewater system” IoT monitoring, “wastewater system monitoring”, “Real-time” monitoring “sewage system”, “Storm Water Management”, “machine learning” sewage systems, “stormwater” in sewage systems, “Infiltration and Inflow” wastewater, “I/I” machine learning.

2.2 Experiment Methodology

2.2.1 Study Area

Kristianstad which is located in the east part of Skåne county in the most southern part of Sweden, it has Sweden’s lowest point that is lower than two meters of the sea level. This entails a higher pressure from the groundwater onto the wastewater channels than what can be found in other regions as well as a bigger risk for flooding from a rising sea level occurring from the continuous climate change. Kristianstad municipality has as prevention matters, built embankments around the city and is using pump stations to direct the water away from the worst risk areas and infrastructure in the town. Kristianstad municipality has also during the last years installed sensors around the town to monitor the situation by analyzing the data which can affect both flooding along with higher tolls on to the

wastewater channels. For this thesis project we had access to all the sensors within Näsby which is the northern area of Kristianstad shown in figure 1.

Figure 1. Map of Näsby [21].

2.2.3 Experiment Research Question

Which data analysis and machine learning techniques are best for detecting leakages in wastewater systems?

2.2.4 Data Collection

In this study we had access to all the installed sensors within the Näsby region, the information that we can obtain with these sensors are precipitation, groundwater levels, and sewage water volumetric flow. All sensors are connected to a database through a wireless network that can be visualized on the IoT data portal provided to us by the municipality. The data can be retrieved through scripts to obtain both historical data and real-time data.

In this study we used the historical data spanning from March the 5th to May the 5th to train our machine learning model, any time period which is within these two dates will be referred to within the training data timespan and any time period after May the 5th will be referred to after the training data time span. The decision for using data from this time period is because of how recent the sensors were installed therefore the data for previous years and months would be unavailable due to how all rows must be filled for the machine learning algorithm to work. When retrieving the data, it comes as different files for each sensor that we pulled data from, to be able to use this data in our machine learning model we had to first index all the values to the same timestamp intervals so that we have a matching index for all the variables. How this was accomplished was because we retrieved the water consumption from the municipality on a per hour basis, we will use this as our index, we then rounded all the timestamps to the nearest hour and picked the highest value for that timestamp. After processing the data in this fashion we then combined all the files together using append on the *Pandas* dataframe and the result is a CVS file with all values indexed in full rows as shown in figure 2. The Python code that was used to accomplish this will be presented in the appendix along with the code used for the rest of this experiment with a written explanation.

Timestamp	Flow L/Sec	water Level Adjusted Zeropoint	Precipitation	water Use Percent
2021-04-29 16:00:00	5.55	3.78	0	3.9
2021-04-29 17:00:00	6.01	3.78	0	5.3
2021-04-29 18:00:00	6.12	3.78	0	5.9
2021-04-29 19:00:00	6.06	3.78	0	6.6
2021-04-29 20:00:00	7.82	3.78	0.1	6.6
2021-04-29 21:00:00	7.22	3.78	0.3	5.3
2021-04-29 22:00:00	6.63	3.78	0.1	4.6
2021-04-29 23:00:00	7.05	3.78	0.2	4.6
2021-04-30 00:00:00	5.96	3.78	0.1	2.6
2021-04-30 01:00:00	5.49	3.78	0.1	1.3
2021-04-30 02:00:00	4.54	3.78	0	0.7

Figure 2. Shows how training data looks after we have processed the data.

In the table shown in figure 2, we can see that the first column represents timestamps which we will be using as an index for our experiment, the second column represents the flow of wastewater within the pipes which is measured by a sensor within the sewage system and is able to measure how much wastewater flows through the pipes in litres per second. The third column represents the water level adjusted to a zeropoint, this zero point is located within the Netherlands and was calibrated to this point by the company installing the groundwater sensors. The fourth column represents the Precipitation of the area, how this is measured is by using an automatic rain gauge that counts how many millimeters of rain is collected every ten minutes, we then choose the highest value for that hour. The last column in this figure is the average water usage over a 24 hour period in percentages, therefore if we would add up the values over 24 hours it would be 100%.

2.2.4 Simple Linear Regression and Pearson's Correlation

The strain on the sewage pipes and the sewage treatment plants comes from the groundwater pushing on the pipes, rainwater infiltrating into the sewage system, and the daily water usage of the region. To model these variables, we will first use a simple linear regression model to see the relationships between the independent variables and the dependent variable.

$$y = A + B(x) \quad (1)$$

Equation one is for a simple linear regression model, with Y being the variable to be calculated which would be the dependent variable, A is the constant or

intercept, B is the coefficient or slope and X is the independent variable [23]. In this study the Y value will be the volumetric flow of wastewater in the pipe and X will be the groundwater level, precipitation, and daily water use. After we have fed the data that we have collected from the sensors, a prognosis can be made to compare to actual data.

$$B = \frac{N \sum(xy) - \sum x \sum y}{N \sum(x^2) - (\sum x)^2} \quad (2)$$

$$A = \frac{\sum y - B \sum x}{N} \quad (3)$$

To calculate the values of A and B the least-squares method is used and its formulas are shown above, the first step is to calculate the coefficient or slope which would be B, as can be seen in the formula, the first step would be for each data point to calculate x^2 and xy, then we sum all of x, y, xy, and x^2. We then insert this into the first formula with N being the number of data points we are calculating. The result of this equation is our B value for our line of regression. We then use the B value that we have just calculated and the sum of y and x in the next formula to calculate A or the constant of our line of regression. This gives us both the A and B values for our simple linear regression model.

To further find the relationship between the variables we can use Pearson's correlation coefficient which is used to find, mathematically, how strong the relationship between the variables is.

2.2.5 Pearson's correlation coefficient

The research team has calculated a correlation between different values using Pearson's correlation coefficient equation. The equation takes two variables and compares these two by drawing an imaginary line in between the values (as accurately as possible) when put on a graph. The coefficient (r) calculated then indicates how well these values "stick together". The coefficient result will always be somewhere from -1 to +1. The strength of the correlation is shown by how far

away from the result being 0 is. If r > .5 the association is considered high, if .5 > r > .3 the association is considered relatively high and if r < .3 the association is considered low. When r is negative the result instead shows a negative correlation in the same "strength scale" as presented above for positive r. That means that the lower r is, the more "anti-correlated" the values are. If the result is 0, there is simply no correlation between the values [22].

Below the equation is presented where x and y represent two different variables. Variables in the equation with an i (for example x_i) stands for the sample of the variables and $\underline{x}$ represents the mean value of those.

$$r = \frac{\sum (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum (x_i - \underline{x})^2 \sum (y_i - \underline{y})^2}} \quad (4)$$

2.2.6 Multi-Linear Regression Machine Learning Model

We will use a multi-linear regression model to map the relationship between the observed sewage pipe volumetric flow and the variables that we have taken into account. If there is a relationship between precipitation, groundwater to the volumetric flow rate this means that rainwater is infiltrating into the sewage system somewhere before that pipe.

$$y = b + a_0 \times x_0 + a_1 \times x_1 + a_2 \times x_2 \ldots \quad (5)$$

With the multi-linear regression model formula shown above, we can see how the predictions are constructed. Y is the variable to be calculated, or the dependent variable, which in this study is the volumetric flow of the sewage pipe in liters per second. X represents the values of the different factors affecting the volumetric flow, the independent variables. A is the model parameters, B being the constant, and A_0 onwards are the coefficient or weight that the variable holds.

2.2.7 Creating and Training the Linear Regression Models

To calculate the simple and multi linear regression model we will use Python, a high-level all-purpose programming language, and Python libraries *Pandas* which was used to create and manipulate the sensor data into useable CSV files, *matplotlib.plot* to visually represent the data, and finally *statsmodel.API* was used to calculate the simple and multi-linear regression model and further used to create the various prognosis and calculate predictions.

In this simple linear regression model, the variables we will use are the groundwater level (measured from a zero point located in the Netherlands), precipitation, and hourly water consumption pattern (which was provided to use from the municipality) to calculate and predict the volumetric flow rate of sewage water within a pipe after which we will analyze each variable and select which ones to use on our multi-linear regression model.

The pipe which we have chosen to do our study on is located on the root of the Näsby region sewage network and will allow us to be able to observe a larger volumetric flow rate and a larger area of rainwater catchment to better observe the effects of precipitation on the system. After we have successfully modeled this larger root pipe we will also do an analysis of a smaller pipe that is located further away from the root, this is to see how differently the variables affect each pipe. The locations of where these sensors for these pipes are located can be found in the figure 3.

Figure 3. Map of Näsby sewage network superimposed over a map of Näsby, the red mark is the location of the larger pipe and the green mark is the location of the smaller pipe [21].

To train our model we will be using historical data of the area spanning two months, first using the simple linear regression models we will create a 24-hour prognosis inside of the training data time span for each variable, using the results from this we will then construct a multi-linear regression model. Using the multi-linear regression model we will then create a 24-hour prognosis inside of the training data timespan using actual measurements from sensors and comparing them to actual sensor data from that pipe. To further analyze the effects of the different factors we will then create more models excluding one of the factors to see how this affects the prediction and then compare it to the three-variable model. We will then make a 24-hour prediction of the volumetric flow rate of the sewage water outside of the training data's timespan using real data and comparing it to the sensor's actual data and do further analysis of what would happen if we exclude one of the variables in the model.

3 Literature Review Results

3.1 Literature Review Results Overview

In this chapter, two main techniques for collecting data within sewage channels which are defined as the image processing technique and the sensor technique is presented in section 3.2 and 3.3. In the related work section of 3.4, different ways of analyzing data regarding inflow and infiltration as well as other studies concerning the topic are presented. In section 3.5 a discussion based on the findings of the literature review can be found and in section 3.6 a short conclusion is presented.

3.2 The Image Processing Technique

Some monitoring techniques that have been deployed in sewage pipes are using a camera and image recognition to detect anomalies which could be objects/items that have in some way entered the pipes (flushed down items of clothes etc.) as well as seeing differences in water flow [9]. By using this type of technique it is possible via image analysis to calculate the flow rate based on water levels within the pipes. This is done using direct visual inspection and recording, digital image processing (DIP), and implemented machine learning, broadly known as deep learning. Direct visual inspection and recording can be explained to be the actual video recordings of the pipe flow. The image in this context often looks blurry, is black and white, and difficult to process for the human eye. Digital image processing (DIP) is then rendering and creates a threshold of the grayscale image. This can be referred to as the creation of the binary map of the image. An image in which the pixels are replaced by 0’s and 1’s to make it possible for a computer to analyze the image based on the binary image. Based on the binary image two boundary lines are created between the wastewater and the pipes. The purpose of this is to find the water level of the pipe. Using deep learning, the program was able to detect which part of the images is water and what is made up by pipe and could thereby split the images into sections. With the sections created an

imaginary circle had to be calculated into the segmented image to determine the size of the pipe as well as inserting the intersection running horizontally through the middle of the circle. Based on the point where the imaginary circle and the boundary line met the water level was obtained. By obtaining the depth of the water in the pipe the water flow rate could be calculated via the values; pipe-width, water-level, roughness coefficiency (based on the material of the pipe), and the length of the pipe. Another image processing setup was used in an experiment conducted in Germany on open channels to measure flow rate. Infrared cameras were set up in different locations along a river as well as an open channel [10]. The goal was to see how well the infrared camera could measure the water flow rate. An Acoustic Doppler Current Profiler (ADCP) was being used as a control measure to compare the results to the infrared cameras. The results of the measures showed that the maximum deviation was 15% between the data that was collected from the two different measurement techniques. The study concluded that using infrared cameras to measure water flow is a very inexpensive way to monitor wastewater. Another positive that was concluded was that no measurement equipment was needed to touch the measured medium.

3.3 The Sensor Network Technique

This technique is based on using different sensors that send the data in real-time to a database where it is stored. The different sensors in the system could be water-level sensors in waste-water pipes, groundwater sensors, temperature sensors, rain-level sensors to track any correlation between I/I and rainfall, and water-level sensors in drainage systems [18]. The data from the water-level sensors in waste-water pipes could be used to calculate the flow rate, the data from the groundwater sensors could be used to calculate the movement of groundwater and potential I/I to the wastewater pipes, the data from temperature sensors could be used to show the impact on drainage system during snowmelt, the data collected from rain-level sensors could help track any correlation between I/I and rainfall and data collected from water-level sensors in drainage systems could help simulate any potential

future flooding [11]. In a study issued in MDPI of 2019, a wireless network monitoring system was built up using ZigBee and 4G to transmit signals in-between sensors and servers [19]. To analyze the incoming data, machine learning algorithms were used to quickly detect any leakages in the monitored pipes. The study concluded that the identification method the research team had built up would work effectively in identifying water pipeline leakages. Another study which was conducted in Lille, France, used hydraulic sensors to measure the water flow in wastewater pipes [12]. The study conducted that the hydraulic sensors gave reliant data with very little maintenance work. The study also conducted the turbidity sensors continuously be objected to dirt from which the sensors needed to be cleaned to function properly.

3.4 Related Work

One study conducted in Uppsala University explored if leaking drinking water channels could have a significant impact on infiltration water in wastewater channels. The study was conducted using different machine learning algorithms which were simple linear regression and multiple linear regression analysis as well as correlation analysis. The research concluded that the linear regression analysis showed a positive correlation between Swedish municipalities' non-billed water and infiltration water entering the wastewater channel significantly [13].

In Hong-Kong where heavy rainfalls can be considered a normal annual happening, a research team conducted a study with the goal of implementing a system to real-time monitor levels in drainage pipes. The experiment was conducted using IoT devices using sensors and machine learning to analyse the flowrate and water level data which was collected. The team concluded that of the two machine learning algorithms used in the project, both Artificial Neural Network (ANN) and cross-validation method showed great potential for implementation in future monitoring systems [14].

In Knivsta which is located between Uppsala and Stockholm in Sweden, the population is quickly rising. More housing is in the time of writing being planned for further development. This is something that would affect the drainage pipes leading to Knivstaåns river and a research team chose to conduct simulations on how this may come to affect the area. The simulation was made using AutoCAD and SWMM, and the methods for analysing the data were made using infiltration methods and runoff-coefficients which are both included in SWMM [15].

A study issued in MDPI in August 2020 had the goal of comparing different AI-based methods for analyzing inflow and infiltration in Sewer Subcatchments. The two methods used in the project were Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multilayer Perceptron Neural Network (MLPNN) using the variables time, rainfall, water consumption, and wastewater flow rate to predict wastewater flow rate at the corresponding time in hours. The research team concluded that both of the developed AI-methods showed the result that rainwater had an instant impact on infiltration to the wastewater channels while the ANFIS method overall showed a higher performance [17].

Though the main conception is that infiltration and inflow water affects the wastewater channels negatively, a literature study was conducted on how infiltration and inflow affect the wastewater channels as well as the positives and negatives of this. The research found that the positives of having I/I in the wastewater channels could be better drainage, easier to control the groundwater levels, less odour and corrosion, higher velocities leads to better self-cleaning and less pollutions in the happening of an overflow in the pipes. The negatives found was; a larger energy consumption, more water makes a larger need for maintenance, larger use of chemicals to “clean” more water, less capacity for future connections, higher potential of flooding and blockages and more water makes the pipes age quicker. The study concludes that though there are positives to a small amount of I/I water, the overall negative effects must be seen as worse [16].

A study conducted on how to improve real-time SWMM flow forecasts using a machine learning approach was issued in EGU in 2020. The approach was to first calibrate an SWMM model using geospatial and hydro-meteorological data and a genetic algorithm (GA). The data was then processed using an artificial neural network to improve the real-time forecasts. The team concluded that some of the GA calibrated data did not show acceptable results on it's own and that using a bias system which also processes the data using ANN will improve the result to an acceptable level [20].

3.5 Literature Discussion

The result in summary from the different ways of monitoring wastewater channels could be separated into "The image processing technique" and "The sensor technique". The obvious positive of using the image processing technique when measuring flow rate is that none of the measurement equipment needs to touch the measured water and therefore not interfere with the flow. However, the result of the image processing technique in most of the cases showed, though it is vague, miscalculated results up to 15%. Based on the fact that the infiltration and inflow rates can be argued as "high" when being around 30% of the total amount of water in the wastewater channel, a 15% miscalculation could be devastating for the long-term measurement effect when trying to optimize the system. Other drawbacks with image processing are the impact of dirt on the lens which could affect the measurements and the need of having the infrared camera completely steady at all times which could be problematic in a wastewater channel. The positives of using "the sensor technique" is that it is standardized and does not have to be calibrated based on the environment it is placed in which makes the implementation very simple. This will affect an area where maybe up to a few hundred measurement points will be implemented. Most of these types of sensors also run on long-lasting batteries as the sensors collect raw data that is then sent directly to a server. The sensors can because of this be placed in remote areas without the need for electricity. A negative regarding the "sensor technique" is

that material can affect the measurement as the sensors often interfere with the corresponding water.

3.6 Conclusion

The related work shows that there are many different ways of both monitoring (gathering data from) sewage networks as well as analyzing the collected data. The academic review that has been conducted also shows that there is a lack of fully implemented monitoring systems in drainage- and wastewater channels as most of the experiments have been conducted on smaller scales for future implementation and that there is much more to do in the field to help optimize infiltration and inflow.

Image processing is something that is still being improved and looking at how cheap the method could be implemented is something to consider in the future. However, development is needed to make it a valid option for implementing in hundreds of places in sewage pipes and it is a long way to go before being considered more effective than a sensor system to measure flow rate through its advantages.

There are different types of machine learning algorithms and AI methods that have been tested and used in the data analysis to find infiltration and inflow water. Most of the experiments found on algorithms to process data collected from sewage systems had combined different algorithms in many cases using a machine-learning algorithm to calibrate the data and then an AI-based method to predict future outcomes. The studies presented in the literature review only using machine learning algorithms showed however that the multiple linear regression and the cross-validation method worked successfully and most efficiently based on the algorithms tested in finding Infiltration and Inflow water.

4. Experiment Result

4.1 Results Overview

Tables one and two contain the values used to create the prognosis and the values of actual volumetric flow and predicted volumetric flow for the larger pipe using all three variables in the model. Figures four to six show scatter diagrams with their respective lines of regression, in the figure description is also the formula for the line of regression and Pearson's correlation coefficient. Figures seven to nine represent the 24-hour prognosis created using the simple linear regression model and its comparison to the actual volumetric flow collected from the sensor. Figures ten to eighteen contains the visual representation of our prediction using the multi-linear regression models, we can observe from the figures that the model has been able to predict the volumetric flow rate to a certain degree but to be able to train a better model we have to further understand the effects of the independent variables to the dependent variable.

All figures will be represented in the appendix while the figures included in the results section will only include the significant figures. These figures will be further compared and discussed within the discussion sections of the degree project.

4.2 Simple Linear Regression and Pearson's Correlation Coefficient

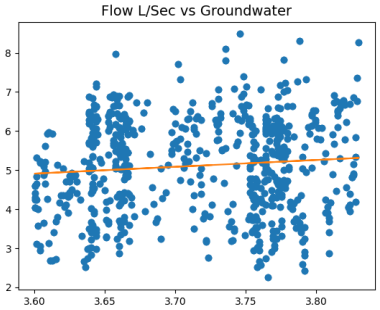

Figure 4. Rate of flow vs groundwater level.

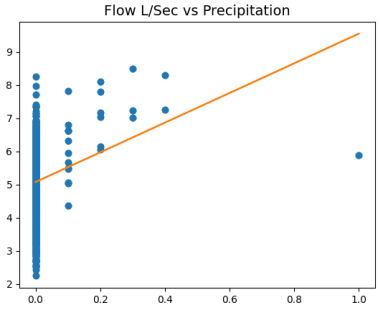

Figure 5. Rate of flow vs precipitation.

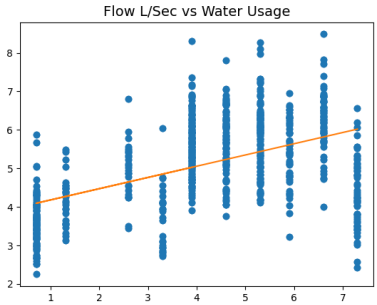

Figure 6. Rate of flow vs water usage.

Figure four represents the relationship of rate of flow and groundwater, this variable resulted in a very low Pearson's correlation coefficient at only 0.094 which is considered a very low to no correlation. Figure five is the relationship between flow rate and precipitation which resulted in a low Pearson's correlation coefficient of 0.202 which is considered a low correlation and finally figure six represents the relationship between rate of flow and water usage which resulted in a Pearson's correlation coefficient of 0.496 which would be considered relatively high.

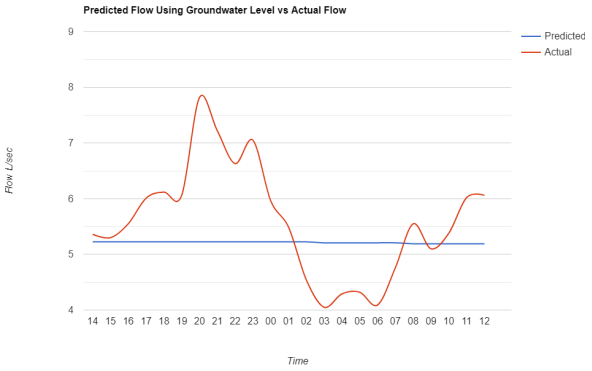

Figure 7. Prognosis using groundwater level.

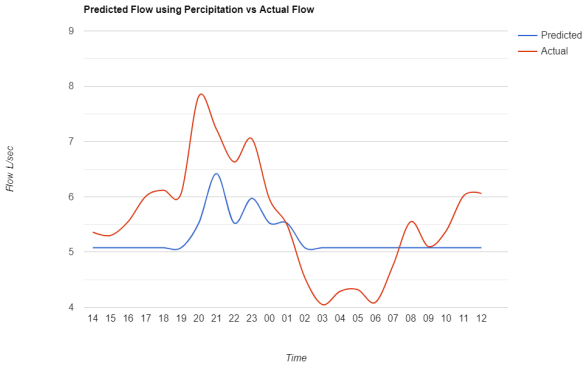

Figure 8. Prognosis using precipitation.

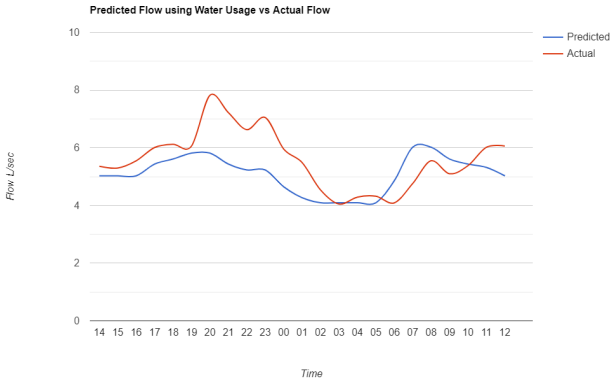

Figure 9. Prognosis using water usage.

Figures seven to nine are the predictions using the single independent variable ibn our simple linear regression model, for figure seven the variable used was groundwater, for figure eight precipitation was used and finally for figure nine only the water usage was used to create the predictions. Using these models we can now predict which multi-linear model will be best to use and compare them to models using different variables.

4.3 Multi-Linear Regression

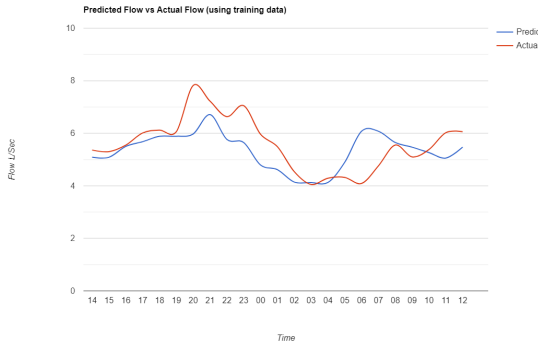

Figure 10. prognosis using all three variables.

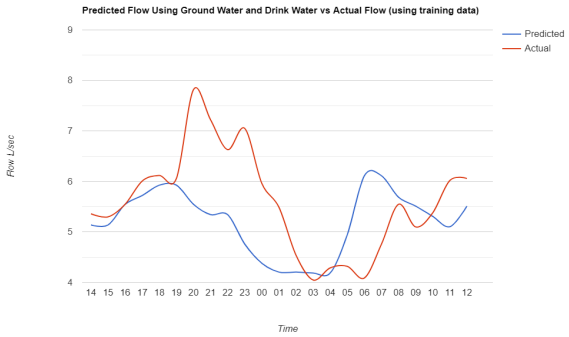

Figure 11. prognosis using groundwater and water usage.

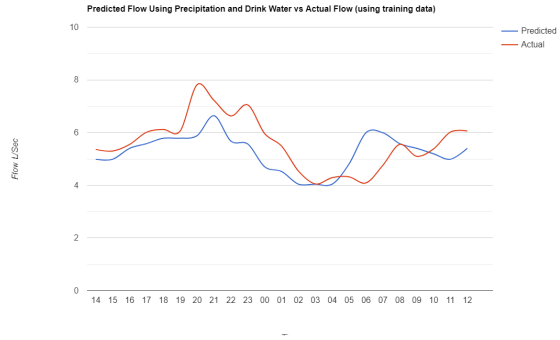

Figure 12. Prognosis using precipitation and water usage.

Figure 13. Prognosis using precipitation and groundwater.

Figures ten to thirteen represents the 24-hour prognosis made using our multi-linear models to see which of our models are best at predicting the volumetric flow rate of the wastewater within the sewage system; the values used to create these predictions can be seen in table one within the appendix. The first figure, figure ten, represents the prognosis made using all three variables which were available. Figure eleven represents the prognosis using only the groundwater and water usage as the independent variables. Figure twelve shows the prognosis using only precipitation and water usage in our model and finally figure thirteen represents the prognosis omitting the water usage variable.

4.4 Discussion

Figure four presented in section 4.2 shows the groundwater compared to flow, observing the graph and the Pearson's correlation coefficient it can be argued that ground would be an inefficient variable to use and will not add much to our models due to the lack of relationship between the two variables. Figure five shows the relationship between rain and flow, here a weak correlation between the variables can be observed and shows that it would be a good variable to use because even if the correlation is weak. Figure 6 shows the water usage relationship to the flow. Here the correlation is much stronger than the other variables and would be a very important variable for us to use in our final model.

Figures seven to nine are the predictions using the simple linear equation which were calculated using the least squares method; these formulas can be observed in the appendix. Figure seven shows the prediction using only groundwater, because of the lack of movement in the predicted curve, it can further be argued that groundwater would be a weak variable to be used in the final model. Figure eight represents the prediction made using precipitation, it can be seen from how the curve is able to predict the same peaks as the actual flow, that it would be a useful variable in the final model if we wish to predict and finally figure 9 shows the prediction using water usage as the variable, because it is able to predict the curve of the line it can be argued that it is a vital variable in our final model.

In section 4.3 can see that the similarities between figures ten and twelve which could indicate that groundwater is not needed in the model due to the lack of change but this could be because the groundwater changes so little throughout the day, for example in the time period of the predictions as seen in table one, the groundwater changed by two centimeters and in our collected data the groundwater changed by 25 centimeters throughout our training data. Figure thirteen is the only graph which excludes the water usage variable, and as predicted, is vital to the prediction of the flow and shows little resemblance to the actual curve except for the peaks caused by rainfall. Figure eleven clearly

demonstrates why even if there is a weak correlation between the variables it is still important to include because of the missing peaks which are seen in every other graph in both the predicted and actual curves.

With these results, a best model for analyzing inflow and leakages can now be constructed which could be argued is the model represented by figure 10, which represents the multi-linear regression model using all three variables. Though this model does contain groundwater which is very weakly correlated to the volumetric flow rate of the sewage water and therefore could be argued to be a bad variable. The variable is still very vital because the groundwater, unlike the precipitation which gradually increases the volumetric flow rate, would act more as a on and off switch because once the groundwater level is above the pipes which contains cracks it would not increase the rate of leak if the water continued rising a few centimeters, therefore it would still be needed to track this variable to be able to detect the inflow and infiltration of groundwater.

4.5 Limitations

The data analysis in the experiment part has been done on collected data from a specific area where sensors are still being implemented. This means that data from the area has not been collected for more than nine months. Though it is not the point of the study, it should be mentioned that the outcome of the current predictions presented in the thesis could be misleading due to the lack of historical data to train the machine learning algorithms, such as how much the groundwater might change throughout the whole year. Another limitation that occurred with the data collected is the lack of some data such as water consumption, for this study we used an average in percent of how much water was used over a twenty-four hour period which helped us create a usable model for us to study but a parameter that would have helped the model generate our prognosis much more accurately would be to have the actual water usage in liters per second and to have sensors similar to the others which would be able to relay real-time data.

5. Conclusion

5.1 Conclusion of Experiment

From the literature review it was concluded that a multi-linear regression model would be the best machine learning technique to use for our degree project experiment. By using data collected from sensors placed in different parts throughout the wastewater network it is possible using multi-linear regression to model the different effects that the various factors have on the volumetric flow of the wastewater pipe and in turn able to study the degree of infiltration and inflow caused by these factors and identify where the infiltration may be located. If the same process of training the data of the multi-linear regression model is used across a large range of wastewater pipes, the location of the infiltration and inflow can, in real-time, be located by following which models have the highest degree of change when looking at the variables of precipitation and groundwater, this can be done as future work for this study.

5.2 Future work

What more can be worked on in the future, related to this study, is to introduce new variables that can further affect the volumetric flow rate, such as temperature which can have a large effect on the infiltration and inflow of rainwater because if the temperature is below zero, it can cause a delayed effect in the relationship between precipitation and the volumetric flow of the pipe until temperatures have risen above zero. Additionally, sensors can be installed to further gather data on different variables such as the actual water usage in an area in liters per second, this could have greatly improved the accuracy of our prognoses. An application can also be built to help identify where areas may have high levels of infiltration and inflow which can help notify workers to do maintenance on the pipes in the section of the network to decrease the amount of rain and groundwater being processed by the wastewater treatment plants helping decrease the cost of maintaining the wastewater system and making it so the pipes carry less overall

water increasing the lifespan and decrease upgrades costs to accommodate the infiltration of rain and groundwater.

References

[1] - Romkey, John. Toast of the IoT: The 1990 Interop Internet Toaster. IEEE Consumer Electronics Magazine. 2017. 6. 116-119. 10.1109/MCE.2016.2614740.

[2] - Why the Internet of Things is called Internet of Things: Definition, history, disambiguation. IoT Analytics; 19th of December 2014. Cited 10th of May 2021. Retrieved from: https://iot-analytics.com/internet-of-things-definition/].

[3] - The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. Cisco IBSG, April 2011. Cited 10 of May 2021. Retrieved from: https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[4] - The IoT Rundown For 2020: Stats, Risks, and Solutions. Security today; 13th of January 2020. Cited 10th of May 2021. Retrieved from: https://securitytoday.com/Articles/2020/01/13/The-IoT-Rundown-for-2020.aspx?Page=4] [Number of internet of things (IoT) connected devices worldwide in 2018, 2025 and 2030. Statista; May 2019. Cited 10th of May 2021. Retrieved from: https://www.statista.com/statistics/802690/worldwide-connected-devices-by-access-technology/]

[5] - Cities and the Innovation Economy: Perceptions of Local Leaders. National League of Cities; October 2017. Cited 10th of May 2021. Retrieved from: https://www.nlc.org/wp-content/uploads/2017/10/NLC_CitiesInnovationEconomy_pages1.pdf

[6] - Nowobilska-Majewska, Elwira & Kotowski, Tomasz & Bugajski, Piotr. Impact of atmospheric precipitation on the volume of wastewater inflowing to the treatment plant in Nowy Targ. E3S Web of Conferences. 2020. 171. 01009. 10.1051/e3sconf/202017101009.

[7] - Eriksson, M. Miljörapport 2020. Avloppsverksamheten Stockholm Vatten och Avfall. 2020. Stockholm Vatten och Avfall AB.

[8] - What is Deep Learning?: Why is Deep Learning Important?. DeepAI. Cited the 10th of May 2021. Retrieved from: https://deepai.org/machine-learning-glossary-and-terms/deep-learning

[9] - Ji H, Yoo S, Lee B-J, Koo D, Kang J-H. Measurement of Wastewater Discharge in Sewer Pipes Using Image Analysis. Water [Internet]. MDPI AG; 2020 Jun 22;12(6):1771. Available from: http://dx.doi.org/10.3390/w12061771

[10] - Hansen, I. et al. “An Innovative Image Processing Method for Flow Measurement in Open Channels and Rivers.” 2017.

[11] - Ades Anspach, C. How Flow Measurement Technology Can Save Water: Monitoring use with flow meters is the key to managing valuable resources. 21st of May 2018. Badger Meter. Cited 10th of May 2021. Retrieved from: https://www.pumpsandsystems.com/how-flow-measurement-technology-can-save-water

[12] - Abbas, O., Abou Rjeily, Y., Sadek, M. and Shahrour, I. (2017), A large-scale experimentation of the smart sewage system. Water and Environment Journal, 31: 515-521. https://doi.org/10.1111/wej.12273

[13] - Ringqvist A. Utläckage från vattennät – en betydande källa till tillskottsvatten i spillvattennät? : Linjär regressionsanalys av VA-data från svenska kommuner [Internet] [Dissertation]. 2021. (UPTEC STS). Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-433773

[14] - K. L. Keung, C. K. M. Lee, K. K. H. Ng and C. K. Yeung, "Smart City Application and Analysis: Real-time Urban Drainage Monitoring by IoT Sensors: A Case Study of Hong Kong," 2018 IEEE International Conference on Industrial

Engineering and Engineering Management (IEEM), 2018, pp. 521-525, doi: 10.1109/IEEM.2018.8607303.

[15] - Boholm P. Bestämning av dagvattenflöden i Knivstaåns avrinningsområde [Internet] [Dissertation]. 2012. (UPTEC W). Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-193230

[16] - Ohlin Saletti, A. Infiltration and inflow to wastewater sewer systems: A literature review on risk management and decision support. Chalmers University of Technology. 2021.

[17] - Zhang Z, Laakso T, Wang Z, Pulkkinen S, Ahopelto S, Virrantaus K, et al. Comparative Study of AI-Based Methods—Application of Analyzing Inflow and Infiltration in Sanitary Sewer Subcatchments. Sustainability [Internet]. MDPI AG; 2020 Aug 3;12(15):6254. Available from: http://dx.doi.org/10.3390/su12156254

[18] - Coles, C. Water Sensors: The IoT Solution. Idtechex; 20th of October 2020. Cited the 16th of May 2021. Retrieved from: https://www.idtechex.com/en/research-article/water-sensors-the-iot-solution/21969

[19] - Liu Y, Ma X, Li Y, Tie Y, Zhang Y, Gao J. Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks. Sensors [Internet]. MDPI AG; 2019 Nov 21;19(23):5086. Available from: http://dx.doi.org/10.3390/s19235086

[20] - Snieder, Shamansouri, Cheng, Ding, Graham, Khan. Improved real-time SWMM flow forecasts using two machine learning approaches. 22nd EGU General Assembly, held online 4-8 May, 2020, id.845.

[21] - Wang, Westlund, Johansson, Lindgren. Smart Sewage Water Management and Data Forecast. April 2021. Cited the 15th of May 2021.

[22] - Pearson's product moment correlation. Statistical tutorials and software guides. Laerd Statistics; 2020. Cited the 10th of May 2020. Retrieved from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[23] - The World Almanac and Book of Facts 1993. Pharos Books; New York 1993. Cited the 17th of May.

[24] - Shaker Life Magazine, Fall 2019. Cited the 30th of May, 2021.

[25] - Inflow and Infiltration. EnviroSight; 2020. Cited the 30th of May, 2021. https://inbound.envirosight.com/inflow-and-infiltration

Appendix

Figures and Tables

Table 1. Values used to predict the flow of water within the training data timespan (prediction using all variables)

Time (Hours)	Water Consumption	Precipitation	Groundwater	Actual Flow (L/Sec)	Predicted Flow (L/Sec)
14	3.9	0.0	3.78	5.36	5.09
15	3.9	0.0	3.78	5.30	5.09
16	5.3	0.0	3.78	5.55	5.50
17	5.9	0.0	3.78	6.01	5.68
18	6.6	0.0	3.78	6.12	5.89
19	6.6	0.0	3.78	6.06	5.89
20	5.4	0.1	3.78	7.82	5.99
21	4.6	0.3	3.78	7.22	6.71
22	4.6	0.1	3.78	6.63	5.77
23	2.6	0.2	3.78	7.05	5.65
00	1.3	0.1	3.78	5.96	4.79
01	0.7	0.1	3.78	5.49	4.61
02	0.7	0.0	3.78	4.54	4.14
03	0.7	0.0	3.78	4.05	4.13
04	0.7	0.0	3.77	4.29	4.13
05	3.3	0.0	3.77	4.32	4.89
06	7.3	0.0	3.77	4.09	6.08
07	7.3	0.0	3.77	4.76	6.08
08	5.9	0.0	3.77	5.55	5.65
09	5.3	0.0	3.76	5.10	5.47
10	4.6	0.0	3.76	5.38	5.26

11	3.9	0.0	3.76	6.02	5.06
12	5.3	0.0	3.76	6.02	5.47

Table 2. Values used to predict wastewater flow after training data span

Time	Water Consumption	Precipitation	Groundwater	Actual Flow	Predicted Flow
14	3.9	0.0	3.73	6.23	5.01
15	3.9	0.0	3.73	6.57	5.01
16	5.3	0.0	3.73	6.73	5.42
17	5.9	0.0	3.73	6.95	5.61
18	6.6	0.0	3.74	7.19	5.82
19	6.6	0.0	3.74	7.33	5.82
20	5.4	0.0	3.74	6.77	5.44
21	4.6	0.0	3.74	6.55	5.24
22	4.6	0.0	3.75	6.12	5.24
23	2.6	0.0	3.75	4.30	4.65
00	1.3	0.0	3.75	3.97	4.27
01	0.7	0.0	3.75	3.56	4.10
02	0.7	0.0	3.75	3.80	4.10
03	0.7	0.0	3.76	3.53	4.10
04	0.7	0.0	3.76	3.32	4.11
05	3.3	0.0	3.76	3.78	4.9
06	7.3	0.1	3.76	4.42	6.06
07	7.3	0.1	3.76	6.38	6.54
08	5.9	0.1	3.76	6.21	6.13
09	5.3	0.1	3.77	5.89	5.95
10	4.6	0.0	3.77	6.14	5.75
11	3.9	0.0	3.77	6.22	5.07
12	5.3	0.0	3.77	6.59	5.50

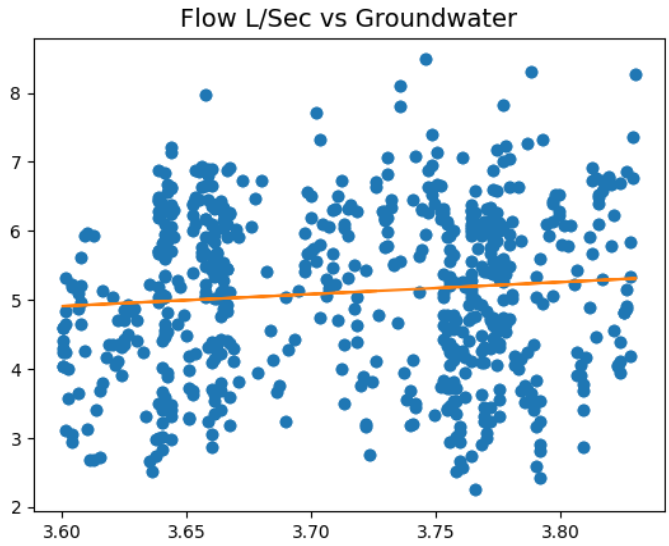

Figure 4. Rate of flow vs groundwater level. Formula: Y = -1.37 + 1.74 ⨯ X Pearson's correlation coefficient: 0.094

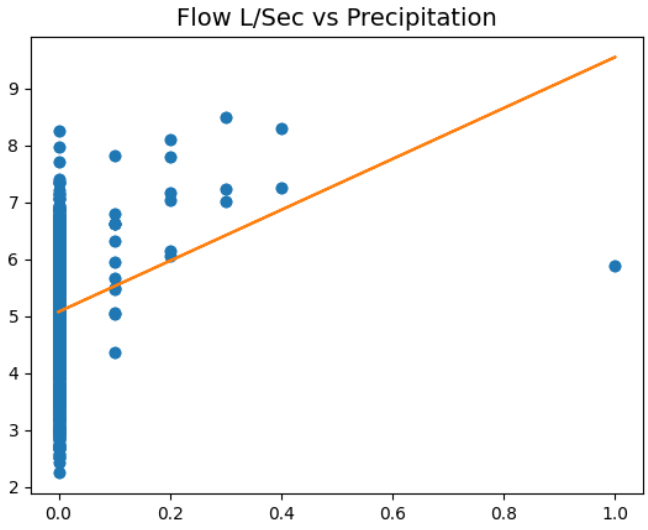

Figure 5. Rate of flow vs precipitation. Formula: Y = 5.077 + 4.47 ⨯ X. Pearson's correlation coefficient: 0.202

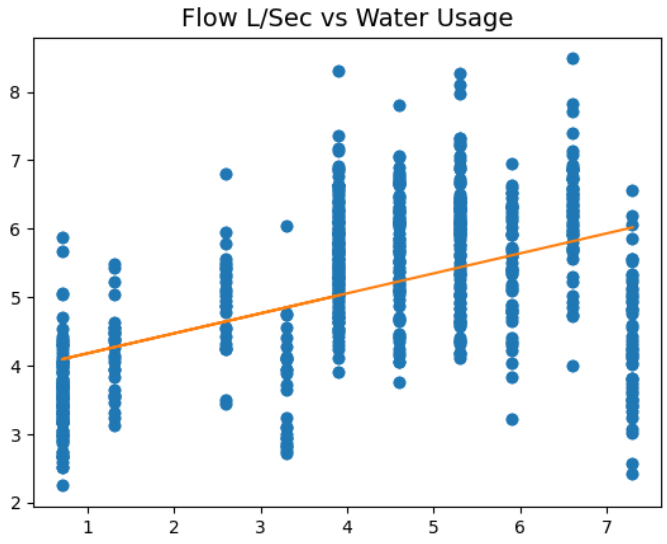

Figure 6. Rate of flow vs water usage. Formula: Y = 3.89 + 0.29 ⨯ X Pearson's correlation coefficient: 0.496

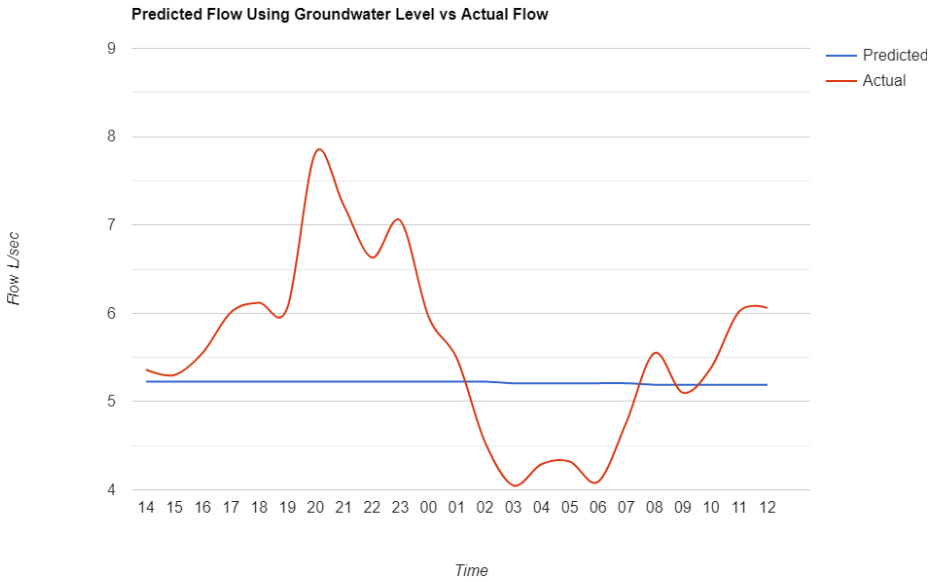

Figure 7. Prognosis using groundwater level.

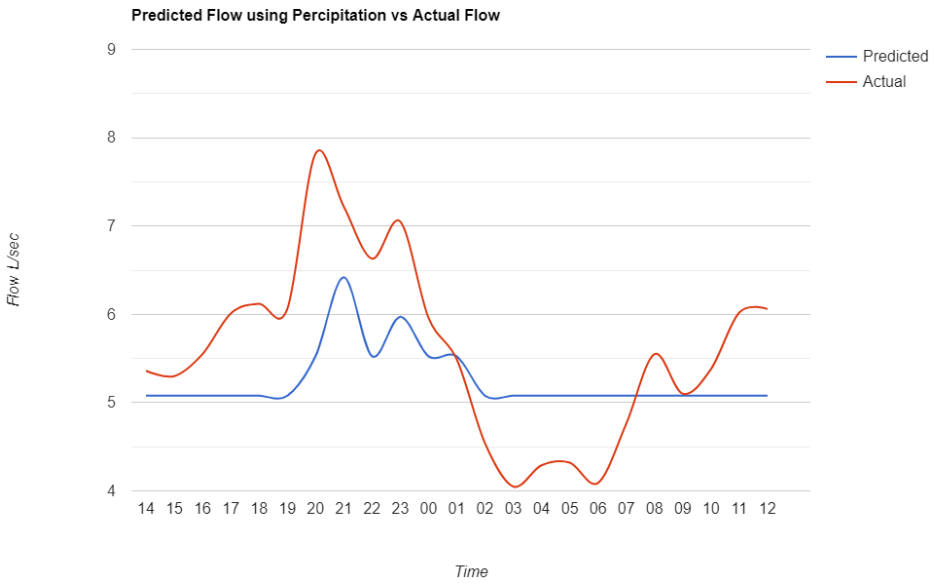

Figure 8. Prognosis using precipitation.

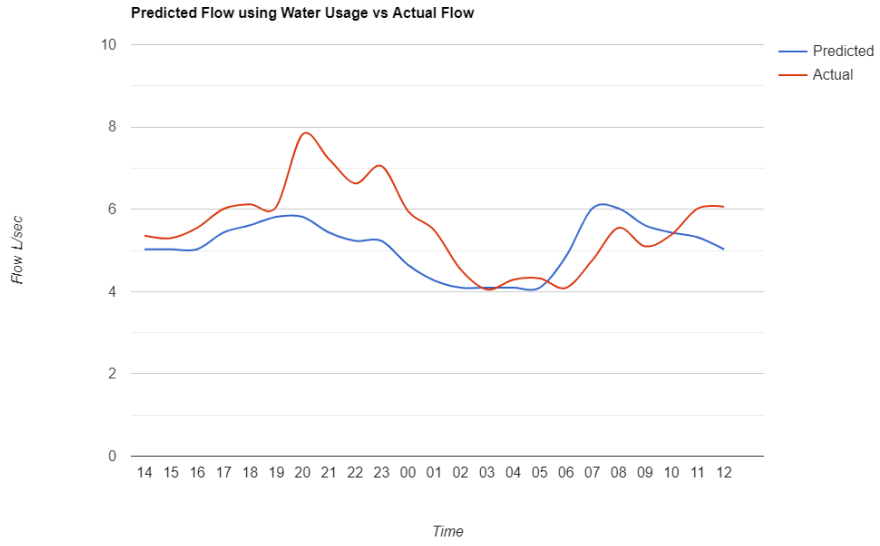

Figure 9. Prognosis using water usage.

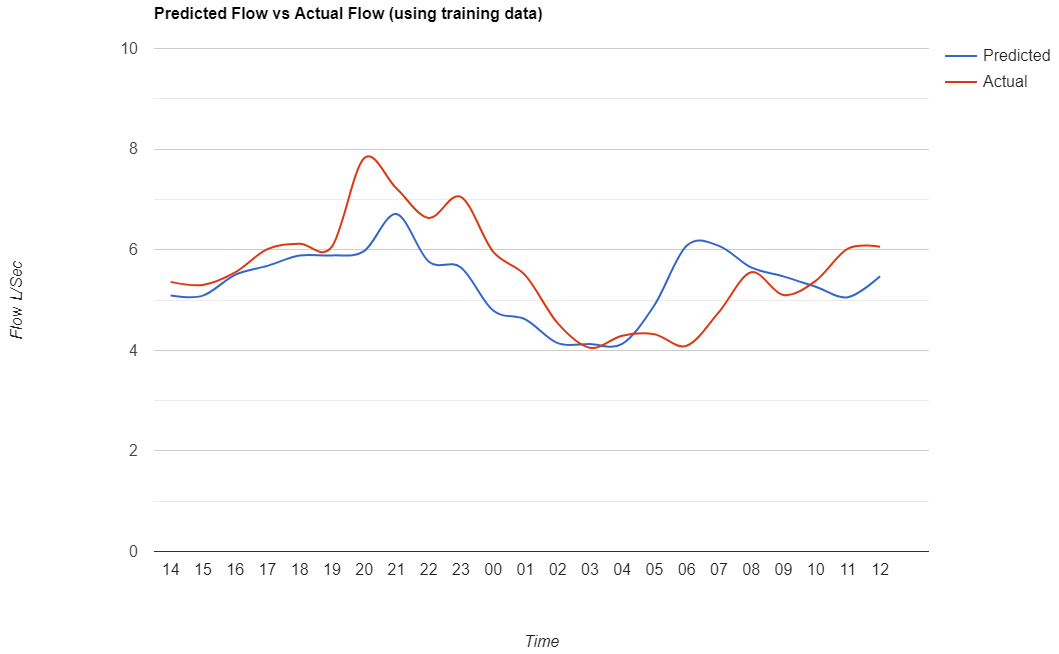

Figure 10. Predicted flow vs actual flow using all three variables within the training data timespan.

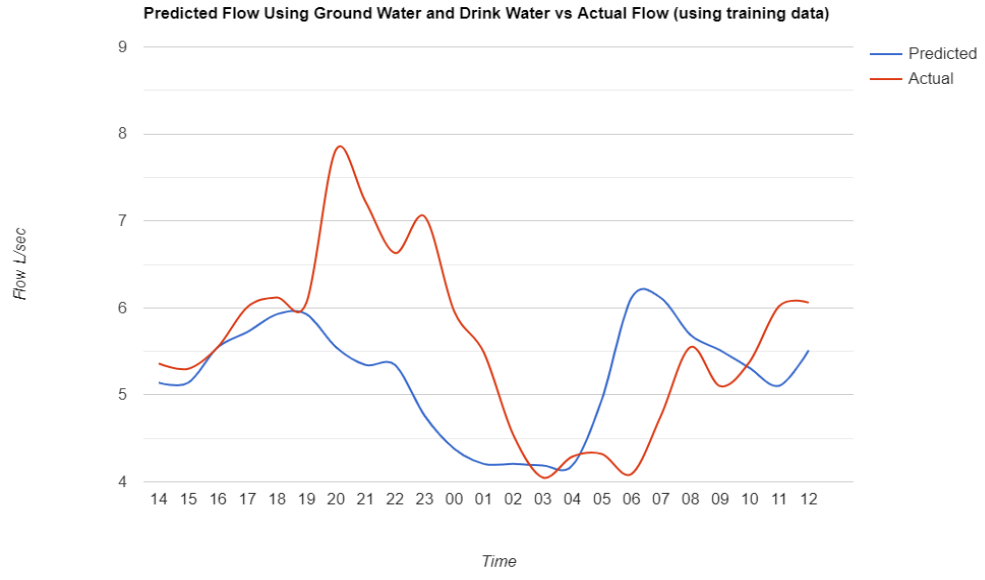

Figure 11. Predicted flow using ground water and water usage within training data timespan.

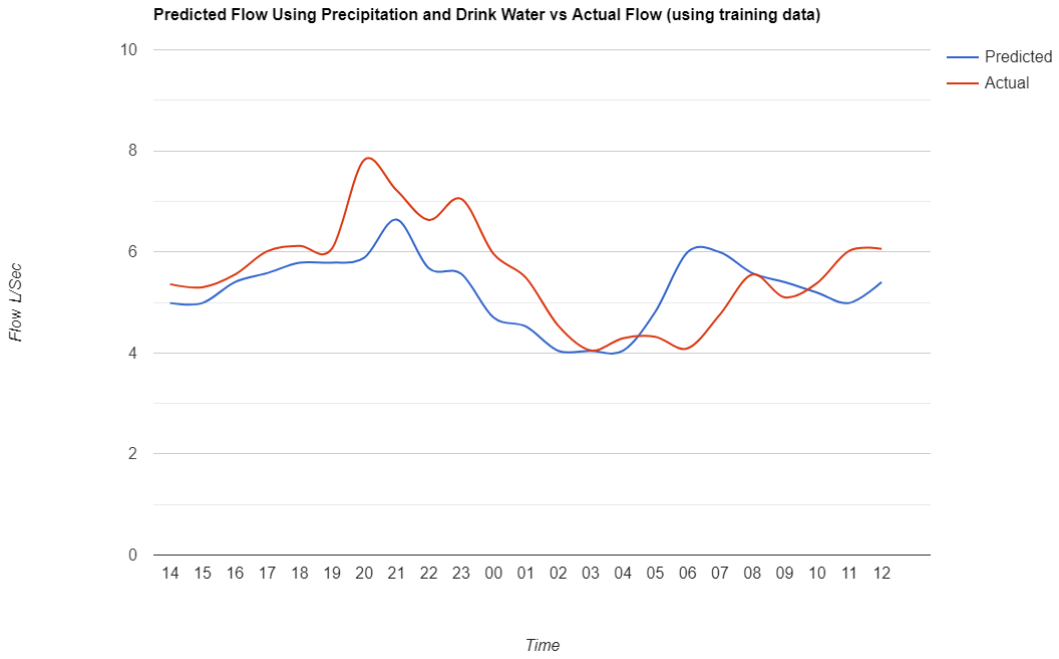

Figure 12. Predicted flow using precipitation and water usage within training model timespan

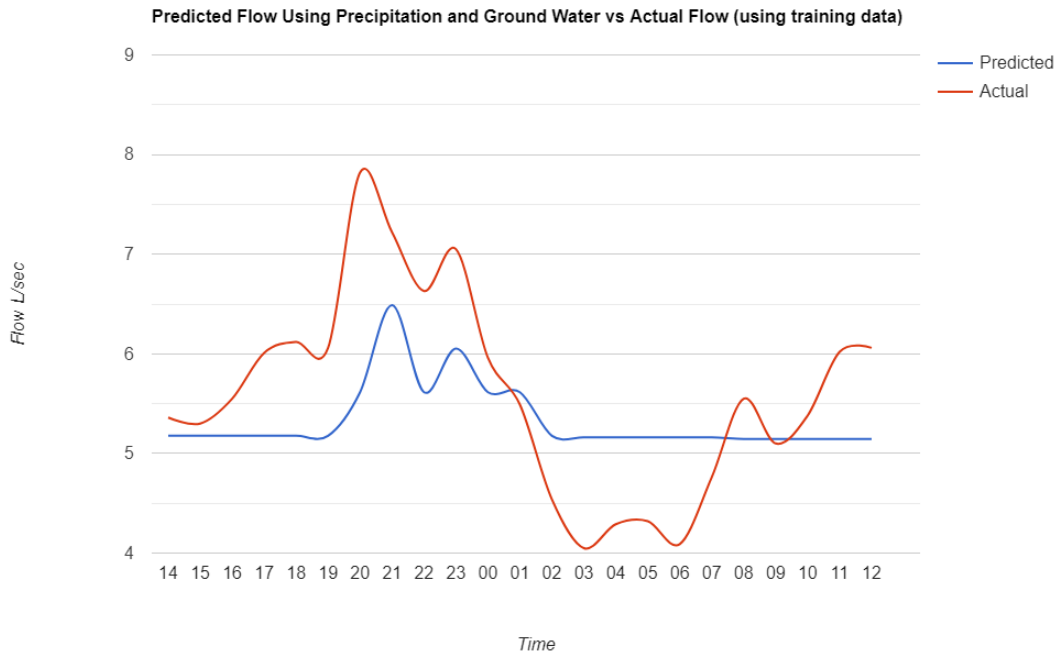

Figure 13. Predicted flow using precipitation and groundwater within training model timespan

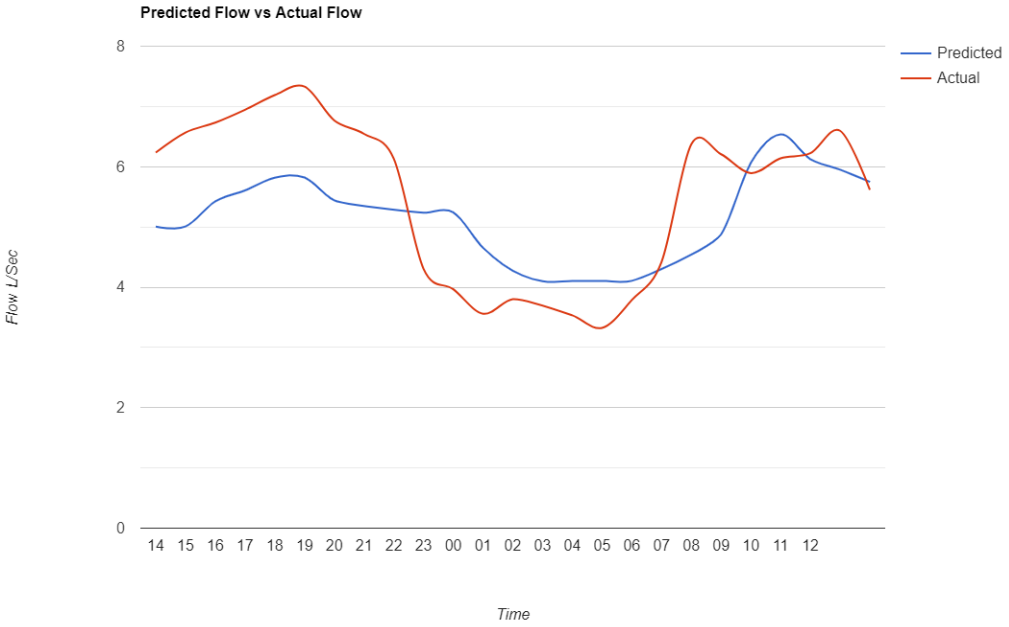

Figure 14. Predicted flow vs actual flow using all three variables after the training data timespan [21].

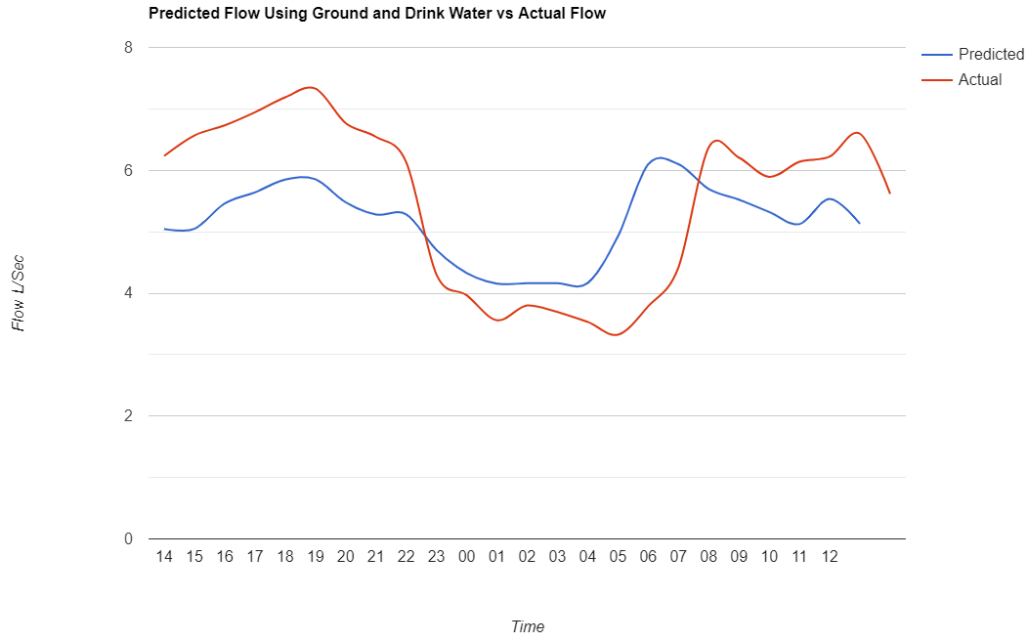

Figure 15. Predicted flow using groundwater and water usage after training data timespan.

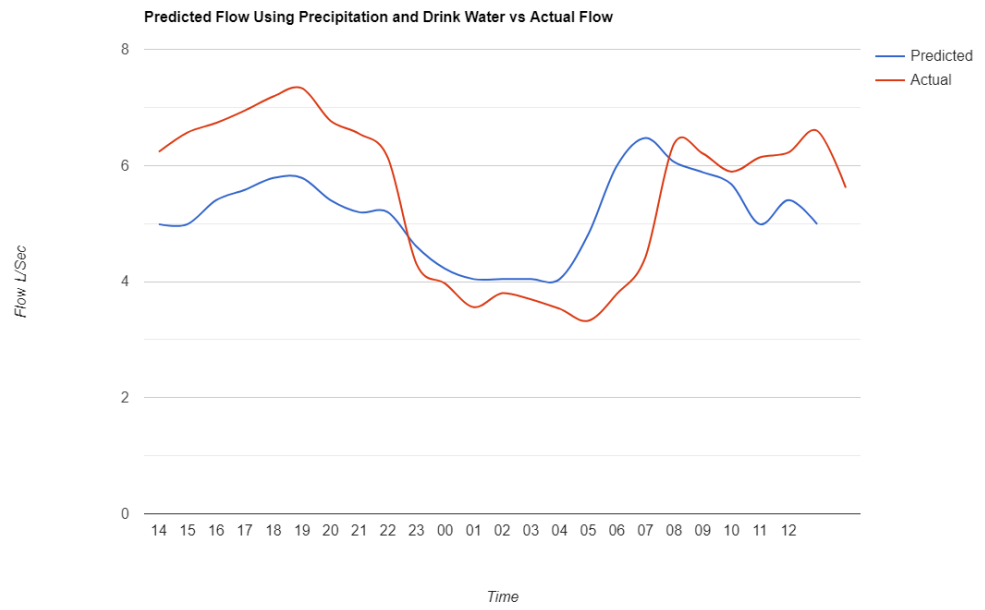

Figure 16. Predicted flow using precipitation and water usage after training data timespan

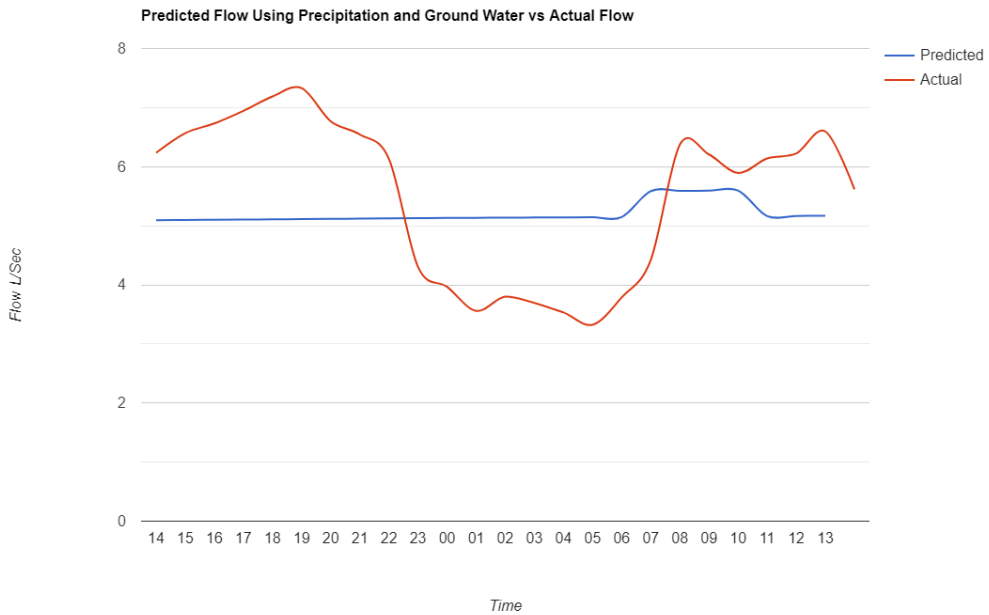

Figure 17. Predicted flow using precipitation and groundwater after training data timespan

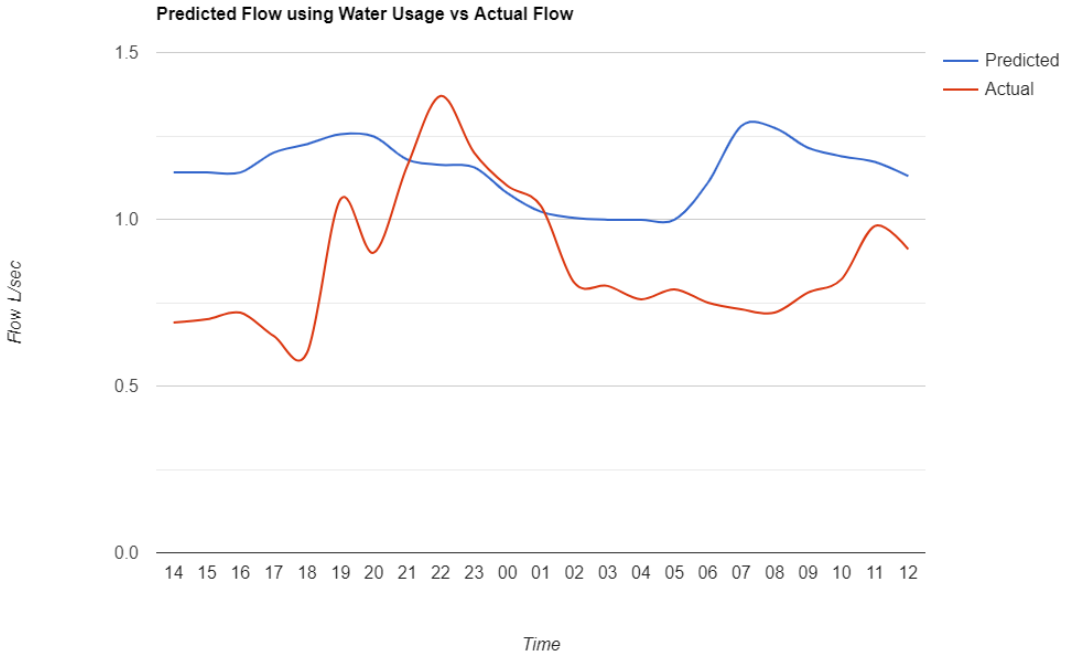

Figure 18. Predicted flow using all three variable on the smaller pipe location

Source Code

Data Collection and Manipulation

```
import pandas as pd
from pandas.io.parsers import read_csv

#array to store the rounded files for combining
li = []

#read the csv file, in this case they used the delimiter ;
flowround = pd.read_csv('flow.csv', delimiter = ";")
#Round each timestamp to the nearest hour
flowround['Timestamp'] = pd.to_datetime(flowround['Timestamp']).round('h')
#save this to a new csv file and set indexing to false
flowround.to_csv("flowRound.csv", index = False)

#read the file again but now set the index to column 0 or in this case the timestamp
flow = pd.read_csv('flowRound.csv', index_col=0)
#pick the highest value for that index
flow = flow[~flow.index.duplicated(keep='first')]
#append to li to get stored
li.append(flow)
#repeat for each varaible
```

After retrieving the data in CSV files either programmatically or from the IoTportal provided, the files are first read and each timestamp rounded to the nearest hour, this is then saved to a new CSV file without indexing and then read again but now setting the index to the timestamp, we then keep only the highest value in that timestamp. this is then appended to the array to later be saved to a file. This is repeated for each variable collected.

```
#turn the array into a frame and save to file.
frame = pd.concat(li, axis = 1)
frame = frame[~frame.index.duplicated(keep='first')]
frame.to_csv("combined.csv")
```

The array with all the variables are transformed into a *pandas* dataframe and checked and deleted any duplicates, keeping the highest value. This is then saved to a CSV file.

Simple Linear Regression and Pearson's Correlation Coefficient

```
1   import pandas as pd
2   import matplotlib.pyplot as plt
3   import numpy as np
4   from sklearn.linear_model import LinearRegression
5   from sklearn import linear_model
6   import statsmodels.api as sm
7   from scipy.stats import pearsonr
8
9   #reads te csv file and sets the index to the Timestamp, any row that has missing data is dropped
10  df = pd.read_csv("combined.csv", index_col='Timestamp', parse_dates=True).dropna()
```

Firstly, we must import all the libraries to be used and then read the csv file setting our index to timestamp and dropping any rows which are missing data because this will cause our machine learning algorithms to fail.

```
75  #linear with statsmodels
76  def simpleLinearRegression(independent):
77      print("\n------------- Simple Linear Regression -------------\n")
78      x = df[[independent]]
79      b = x.values.reshape((-1,1))
80      x = pd.DataFrame(b)
81      y = df['Flow L/Sec']
82      model = LinearRegression().fit(x, y)
83      r_sq = model.score(x, y)
84      print('coefficient of determination:', r_sq)
85      print('intercept:', model.intercept_)
86      print('slope:', model.coef_)
87      x1 = np.array(df[independent])
88      corr, _ = pearsonr(x1, y)
89      print('Pearsons correlation: %.3f' % corr)
```

Here is the method used for the simple linear regression model, firstly we must load all the variables to be used in the regression model which would be an independent variable and the dependent variable being flow. We then fit these variables to the linear regression model and then print the data we want to retrieve. To calculate the Pearson's correlation coefficient, we first have to transform the independent dataframe to an array and then pass the variables to the method provided from our libraries.

Multi-Linear Regression and Predictions

```
63   # multi linear with statsmodels
64   def multiLinearRegression(independent):
65       #change the variables in here to change the model
66       print("\n------------- Multi Linear Regression -------------\n")
67       x = df[independent]
68       y = df['Flow L/Sec']
69       x = sm.add_constant(x)
70       model = sm.OLS(y, x).fit()
71       print_model = model.summary()
72       print(print_model)
```

For the multi-linear regression method, the variables first have to be defined and then a constant is added to the independent variables, we then pass this through our model provided by the imported libraries and then a summary can be printed.

```
91   #Uncomment in predictions() to predictic using those variables
92   def predictions(independent):
93       print("\n------------- Predictions -------------\n")
94       x = df[independent]
95       #If only using one varaible uncomment
96       #b = x.values.reshape((-1,1))
97       #x = pd.DataFrame(b)
98       y = df['Flow L/Sec']
99       lm = LinearRegression()
100      regr = linear_model.LinearRegression()
101      regr.fit(x, y)
102      print('Intercept: \n', regr.intercept_)
103      print('Coefficients: \n', regr.coef_)
104      print ('Predicted Water Flow: \n')
105      #Uncomment according to varaibles choosen
106      for i, l, j, k in zip(waterlevel, Percipitation, waterUse, actualFlow):
107          print(regr.predict([[i, l, j]]))
108          #print(regr.predict([[l, j]]))
109          #print(regr.predict([[i, j]]))
110          #print(regr.predict([[i, l]]))
111          #print(regr.predict([[i]]))
112          #print(regr.predict([[l]]))
113          #print(regr.predict([[j]]))
114          #print(regr.predict([[k]]))
```

This is the method used to create our predictions, the variables are first read and passed to the method imported. this can then be used to print the intercepts and coefficients for each independent variable, finally we can print our predictions using the methods imported. it is important to uncomment according to which variables are used for the prediction. if only one variable is used then it must be

reshaped to be able to be used in the model. the values used for the prediction was hard coded into the program.

```
131  #Must uncomment to match varaibles in predictions function keep inside [] if more than 1 varaible passed
132  #Cannot take a 1D array, so if only passing one varaible to predict with, reshape array in the funtion
133  predictions(['waterLevelAdjustedZeropoint', 'Precipitation', 'waterUsePercent'])
134
135  #independent variables = 'waterLevelAdjustedZeropoint', 'Precipitation', 'waterUsePercent'
136  #keep the varaibles inside []
137  multiLinearRegression(['waterLevelAdjustedZeropoint', 'Precipitation', 'waterUsePercent'])
138
139  #can only take on varaible at a time
140  simpleLinearRegression('Precipitation')
```

This is an example of how these methods are used, only prediction methods must be changed according to which independent variables are used.