I-and-I Writeup

Paul Barton and Brooks Emerick

July 2021

1 Introduction

The cost of wastewater treatment represents a significant component of any municipal budget. The accidental processing of rainwater requires energy, labor, and materials, placing additional financial strain on local governments. The two primary means by which rainwater enters wastewater infrastructure are known as "inflow" and "infiltration." Inflow is the rainwater that rapidly enters the sewage system through properly functioning infrastructure such as drains, manhole covers, downspouts, sump pumps or other rainwater diversion methods. Infiltration represents groundwater which seeps into the sewage system through damaged or eroding infrastructure. The purpose of past research was to estimate the amount of inflow and infiltration, which we will refer to as I.&I., for the borough of Kutztown, as well as to assess the amount of money being lost by processing this excess rainwater in the system. This was carried out by examining the situation at four pump stations which serve as intermediaries connecting properties with the wastewater treatment plant. These four pump stations are referred to as "Briar Cliff," "College Gardens," "Highland," and "Hilltop." Repair work was carried out and completed in October of 2020 for two of these pump stations, Briar Cliff and College Gardens. The repair work (also referred to as "rehabilitation," or "rehab" for short) consisted of procedures to seal leaks in the pipe systems in order to reduce the amount of infiltration. The purpose of this current research is to assess the impact of this rehabilitation as well as to examine the possible financial benefits of future repair work at other sites. Each pump station is equipped with meters which measure the length of time for which the pumps are running each day, and this length of time is recorded. Each pump station also pumps a certain amount each hour (this amount was measured by Vanessa Maybruck, Robert Gould, and Alex...), and so multiplying this constant by the number of hours results in the total amount of wastewater discharged by the station that day, in gallons. These discharge totals are the key to our analysis. We can rate the effectiveness of the repairs by comparing pre-rehab and post-rehab discharge amounts, while taking other factors into account of course.

In this study, we examine the available pump discharge data from January of 2018 through July of 2021 and use statistical methods to evaluate the impact

of the rehabilitation efforts at the College Gardens and Briar Cliff pump stations. In Section 2, we perform a regression analysis on the discharge amounts at each pump station for 2018 and 2019, as well as the months after rehab in 2020 and 2021. In Section 3, we use the household consumption data provided by the Borough to assess the I & I amount as a percentage of total pump discharge. In Section 4, we employ hypothesis testing to determine whether the reduction in discharge is a result of the rehab, or or instead represents a random fluctuation. In addition, we attempt to account for changes in precipitation levels from the pre-rehab and post-rehab time periods.

2 Regression Analysis of Global Data Sets

Our first task was to confirm the extent to which pump discharge is correlated with rainfall events. From a purely visual standpoint, the time series graphs, one for each pump station in 2018 to 2019, would indicate that high pump output coincides with both a high daily precipitation and a higher amount of ground saturation, as shown in Figure 1. The correlation is not as obvious in the case of the time series plots for 2020 to 2021 (Figure 3). Hilltop station, our control (Figures 2 and 4), displayed no noticeable correlation. Ground saturation represents the accumulated precipitation over a two week period. The scatter plots displayed in Figures 5 and 6 are more difficult to interpret, so we plotted linear regression lines for each scatter plot and computed the associated correlation coefficients.

Pump Stations	2018-2019 R^2	2020-2021 R^2
Briar Cliff	0.39419	0.01360
College Gardens	0.28749	0.00158
Highland	0.11811	0.00029
Hilltop	0.00092	0.00267

Table 1: R^2 values measuring the correlation between ground saturation and pump discharge for the years 2018-2019 and the period from December 2020 through July 2021

Our assumption is that a decrease in the R^2 values for a given pump station after the rehabilitation process would possibly indicate that repair work did in fact reduce the amount of rainwater entering the system after precipitation events. If this assumption is correct, then the R^2 values shown in Table 6 do in fact indicate that the rehabilitation was successful for the Briar Cliff and College Gardens pump stations. This would explain why the precipitation and ground saturation time series plots for 2020 to 2021 (Figure 3) seem not to coincide with the pump discharge for those two stations over the given time period, in comparison with the 2018 to 2019 data in Figure . However, while the R^2 values do indicate a correlation for Briar Cliff and College Gardens in

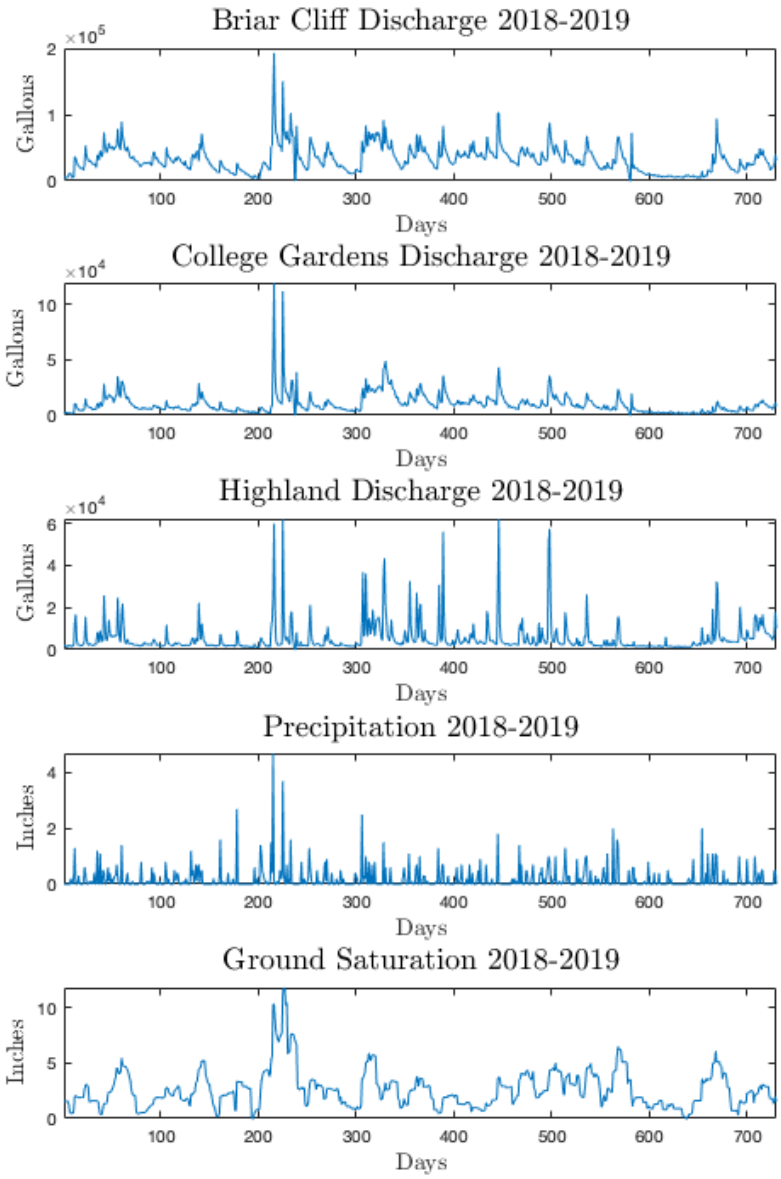

Figure 1: Time Series plots for Briar Cliff, College Gardens and Highland pump stations, 2018-2019

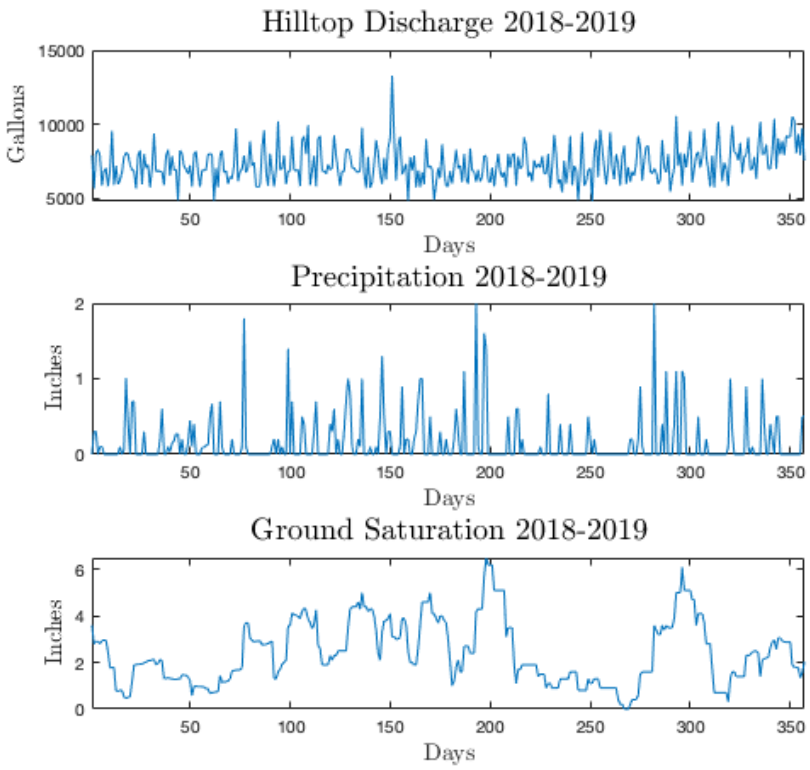

Figure 2: Time Series plots for Hilltop pump station, 2018-2019

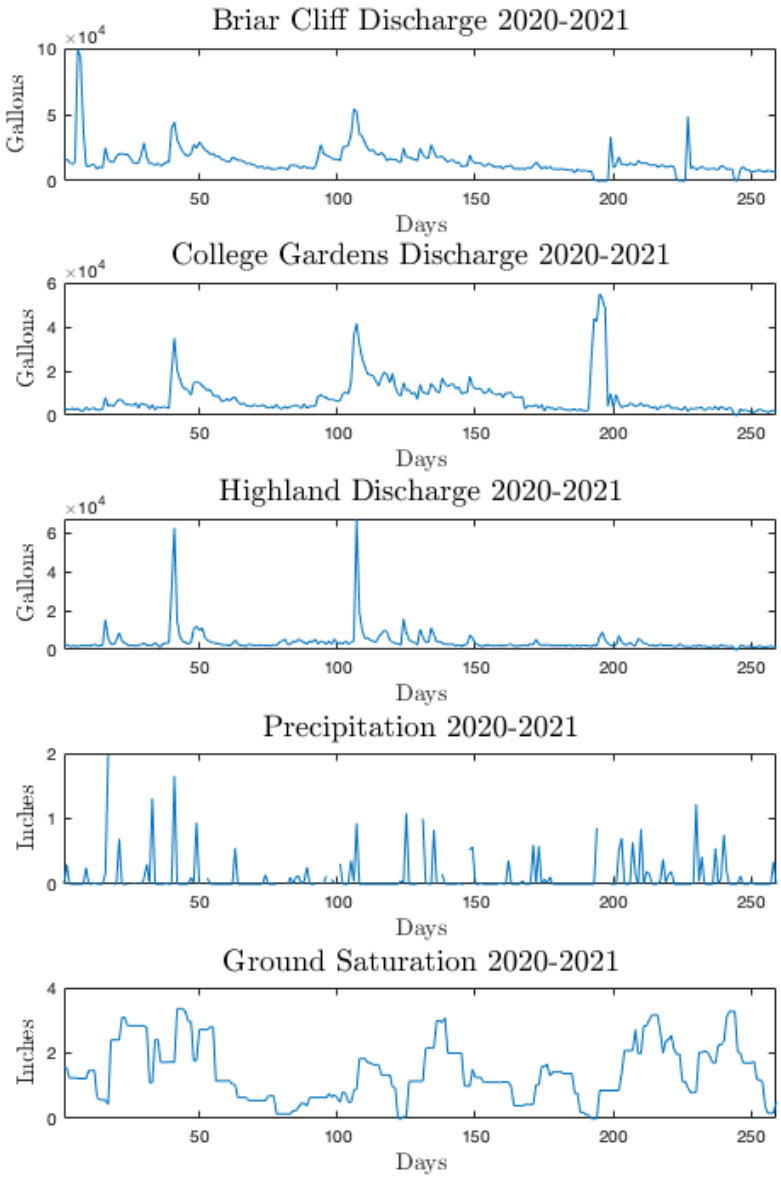

Figure 3: Time Series plots for Briar Cliff, College Gardens and Highland pump stations, 2020-2021

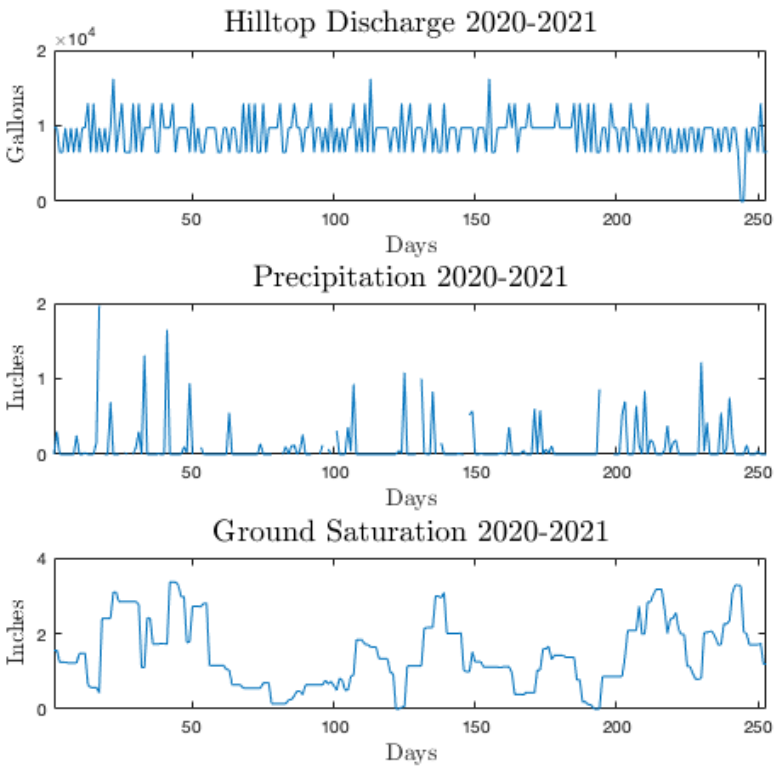

Figure 4: Time Series plots for Hilltop pump station, 2020-2021

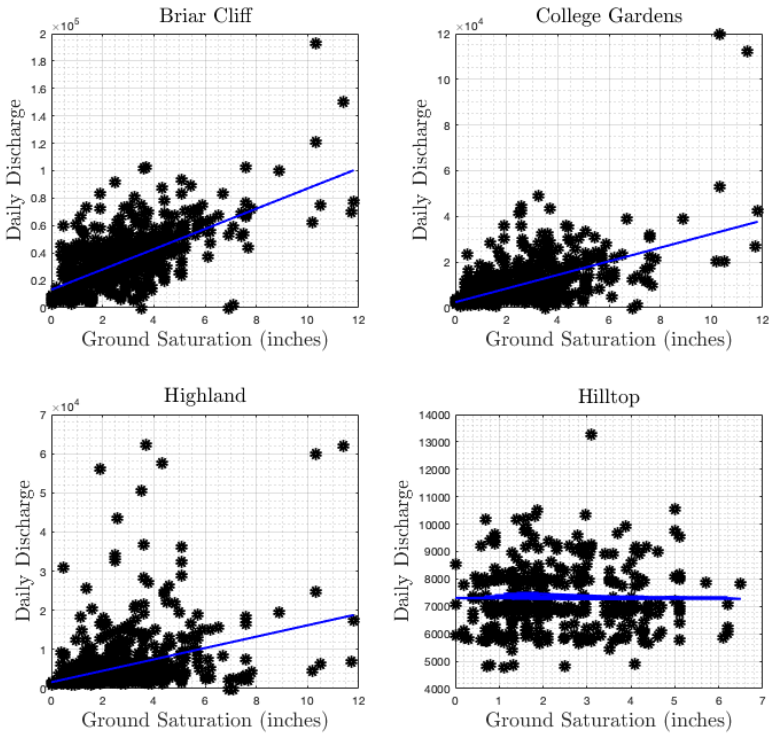

Figure 5: Scatterplot of Discharge VS Ground Saturation, 2018-2019 Change the dots to little circles do this maybe by changing the size?

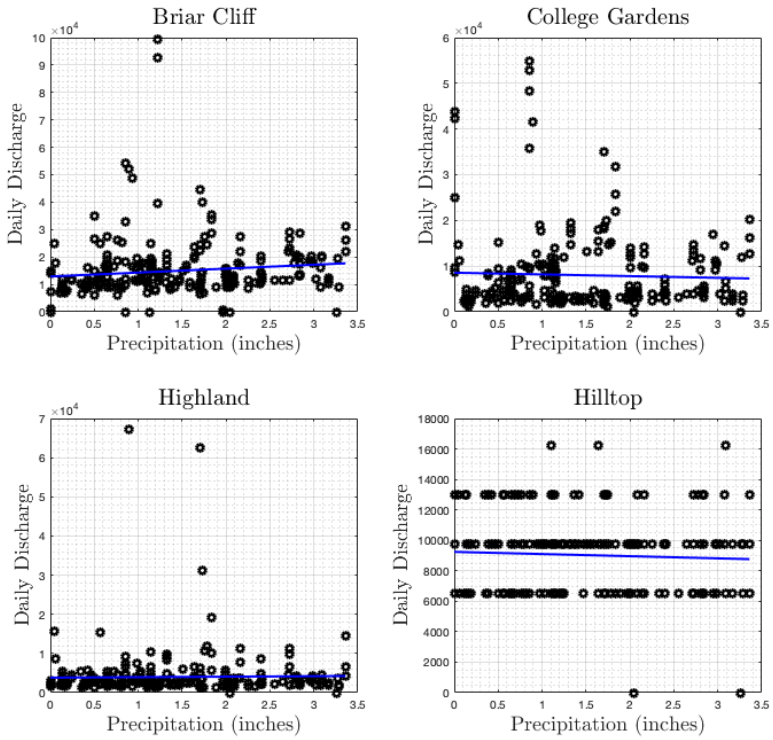

Figure 6: Scatterplot of Discharge VS Ground Saturation, 2020-2021 Change the plots to ground saturation instead of precipitation.

2018 and 2019, they are not particularly high. The definite lack of correlation in the case of the Highland pump station is concerning.

In general, the nature of the data itself seems to obscure the mathematical correlation that appears obvious upon visual examination of the time series graphs. This is most likely because of the high variability in the physical process of precipitation entering the system. Some rain events are brief and intense, others steady but prolonged. A brief, intense event, like a strong thunderstorm lasting a few hours, can trigger a high pump discharge, but not immediately; the pump discharge may be delayed by the time it takes the rain to seep into the system (infiltration); inflow may increase as a result of runoff and sump pumps, but even this can be delayed, or perhaps negligible. In contrast, days or weeks of moderate rainfall can eventually cause a spike in pump discharge as the ground reaches a certain saturation point. This is why we consider two week ground saturation as opposed to simply using daily precipitation totals.

It was noted that in the case of the Highland pump station, some of the pipe network was up to twenty feet below ground, compared to approximately six feet for the Briar Cliff and College Gardens stations. This could cause a longer delay between the rain event and pump discharge as the groundwater has much further to descend before reaching the infrastructure. We conducted an experiment to test this by introducing a delay into the precipitation time series graph and calculating the correlation coefficients for various delay times. The correlation coefficient increased to [...] as we increased the delay to [...]. From this we concluded that following a strong precipitation event there is a delay of approximately [need number here] days until a corresponding high pump discharge. [Perhaps even have a reference to a figure here.]

Another major factor in the relationship between precipitation and pump discharge is cold weather, specifically freezing temperatures. Snow and ice will obviously need to melt before the water enters the groundwater system, and the frozen ground in winter slows or prevents saturation and the infiltration. Eventually the snow and ground melts and the result is a spike in pump discharge, especially if accompanied by heavy rain events (a typical situation in March or April). This pattern further complicates the attempt to find a mathematical correlation between precipitation and discharge. Therefore we conjectured that removing the winter months from the data set could result in a stronger correlation; this was found to not be the case. Overall, the physical mechanisms by which rainwater enters the sewage system are complex and difficult to isolate, involving many factors, some of which we have not even considered in our study, such as topography, soil type and surface runoff.

3 Calculation of I.&I.

It is estimated that approximately 85 percent of water used by residents enters the wastewater system. The remaining 15 percent is lost to boiling water for cooking, drinking water, watering plants, etc. [**?**] [You should cite the source]

We calculate the I.&I. for the region served by each pump station by subtracting the total amount consumed (times 85 percent) from the amount discharged by that pump station in the given time period. We were provided with consumption data in the form of average monthly consumption for the majority of the residences served by the four pump stations. These amounts for each residence were then averaged to give us a monthly average consumption for the pump station. Working with average monthly consumption amounts, which are just the yearly total usage divided by 12, is not as accurate as having a specific average amount for each month, as there could be fluctuations in consumption patterns throughout the year, but it is accurate enough for the purposes of calculating I.&I. for a year. Specifically, we are comparing the pre-rehab I.&I. for the Briar Cliff and College Gardens stations with their post-rehab data. This comparison will also be made purely in terms of pump discharge; calculating the specific I.&I. will help us evaluate the dollar amounts involved.

Pump Stations	I&I 2018-2019	I&I 2020-2021
Briar Cliff	735527.48	207482.79
Percent	66.3%	37.3%
College Gardens	264541.7	191015.5
Percent	77.93	73.1

Table 2: Comparison of pre-rehab and post-rehab monthly average I&I amounts for Briar Cliff and College Gardens pump stations, and corresponding percentages

Because the I&I is calculated by subtracting a constant value each month, we get the same results as we would by working with only pump discharge, but on a different scale. However, the ratio of I&I to total pump discharge is important as well.

4 Assessing Effectiveness of Rehabilitation

Rehabilitation for the Briar Cliff and College Gardens pump stations was completed in October of 2020. The reduction in R^2 values for the Briar Cliff and College Gardens stations from November of 2020 onward may indicate that the repairs were successful, but further procedures were required to confirm this.

4.1 Hypothesis Testing of Discharge

In order to evaluate the effectiveness of the rehab, we conducted hypothesis tests comparing the average daily discharges before and after the rehab. The purpose was to determine whether the rehab reduced the amount of average daily discharge; from this we would be able to assess the reduction in I.&I., if

any. Of course, our visual analysis points very strongly to the fact that heavy precipitation and ground saturation lead to the elevated I.&I., but this would increase the daily average discharge for the entire year. Therefore, successful rehab should in theory lower the daily average discharge.

Our initial test was conducted from the naive standpoint that precipitation is steady from year to year with negligible seasonal fluctuations. Also, by dealing with discharge instead of I.&I. amounts, we are ignoring the impact of resident consumption patterns which may in fact change dramatically over time. The results of this first t-test indicate that for both the Briar Cliff and College Gardens pump stations, the effect of the rehab was to significantly reduce the average daily discharge, as shown in ...

Pump Station	Average Discharge		Hypothesis Test	
	2018-2019	2020-2021	Reject Null	P-Value
Briar Cliff	32380	15095	Y	1.76×10^{-38}
College Gardens	10208	7803	Y	0.0001

Table 3: Comparison of pre-rehab discharge and post-rehab average discharge amounts with corresponding hypothesis test results

Our next test was designed to account for possible seasonal fluctuations in climate and water consumption. For this we look at each month from November 2018 to June of 2019 and individually test daily discharge of these months against the corresponding post-rehab month in the period from November 2020 to June 2021. To be specific, we compared daily discharge in November of 2018 to daily discharge in November of 2020, December of 2018 to December of 2020, and so on up until July of 2019 compared with July of 2021. The results are displayed in ... For each pump station we have a plot of the mean discharge and corresponding standard deviation, blue for 2020 through 2021 data and black for the 2018 through 2019 data. We also have a visual of the results of the test, a dot on the '1' line indicating a significant decrease in mean discharge for that month and a dot on the zero line indicating a lack of significant decrease. For Briar Cliff, we can see a reduced average discharge for every month as evidenced by the row of dots on the '1' line. College Gardens shows a significant decrease for the months November through February, as well as June and July, but not for March, April or May.

4.2 Discharge during Rain Events

The reduced precipitation levels from November 2020 up to and including July 2021 compared to the years 2018 and 2019 make it difficult to accurately assess the rehabilitation efforts. To account for this difference, we first defined a rain event as the union of the days with precipitation greater than or equal to one inch and the days with ground saturation greater than three inches. Next, we

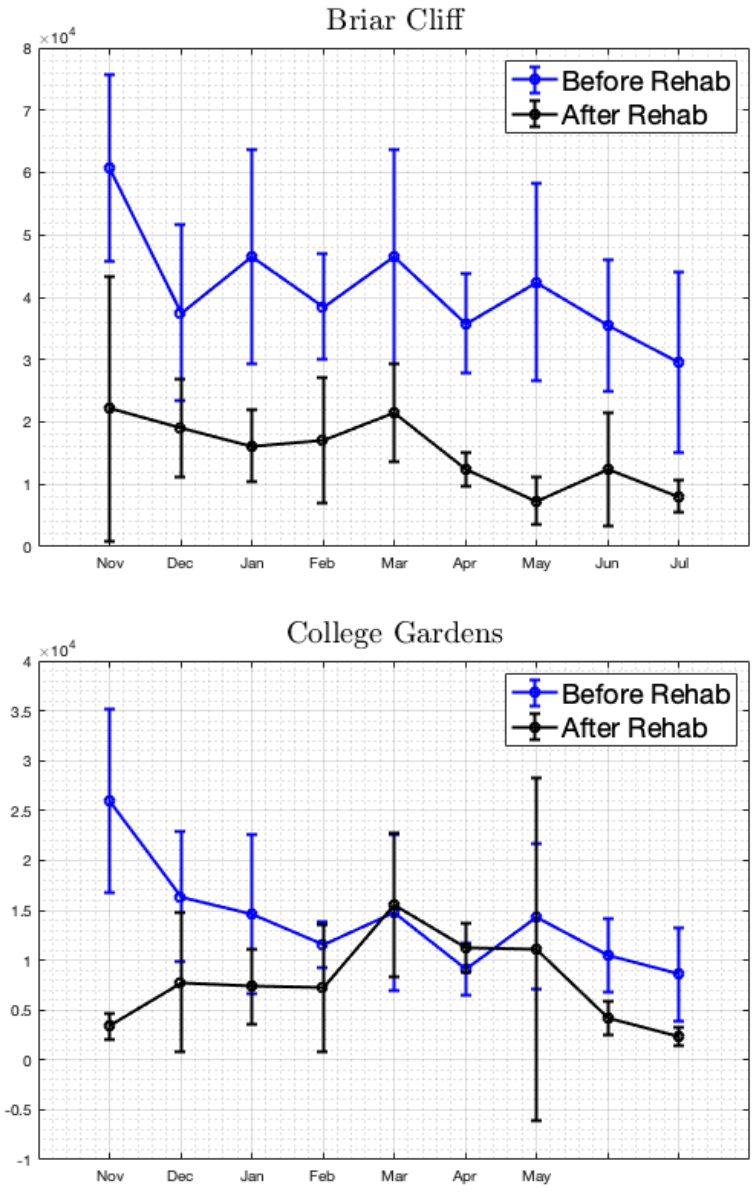

Figure 7: This figure shows the mean discharge and standard deviation for [...]

Month	2018-2019	2020-2021	Reject Null	P-Value
November	60721	22114	Y	1.75×10^{-11}
December	37513	19115	Y	1.72×10^{-8}
January	46443	16123	Y	1.65×10^{-13}
February	38527	17026	Y	4.96×10^{-12}
March	46586	21467	Y	2.32×10^{-10}
April	45748	12424	Y	3.47×10^{-22}
May	42383	7334	Y	8.15×10^{-18}
June	35579	12418	Y	4.88×10^{-13}
July	29650	8033	Y	1.11×10^{-11}
November	26023	3345	Y	1.23×10^{-19}
December	16366	7743	Y	2.33×10^{-6}
January	14600	7338	Y	1.12×10^{-5}
February	11561	7241	Y	0.0006
March	14807	15509	N	0.6425
April	9139	11208	N	0.9988
May	14370	11094	N	0.1670
June	10458	4202	Y	4.98×10^{-12}
July	8586	2370	Y	5.71×10^{-10}

Table 4: Month to month comparison of pre-rehab discharge and post-rehab discharge amounts with corresponding hypothesis test results

removed those days for which there was no rain event from the data and performed the hypothesis tests as before. This way the tests would be comparing mean discharge only for the time periods with precipitation events. Again, the null hypothesis was rejected for both pump stations, with a 95 percent confidence level, indicating a significant decrease in the mean discharge following repair work.

Pump Station	Average Discharge		Hypothesis Test	
	2018-2019	2020-2021	Reject Null	P-Value
Briar Cliff	45615	16381	Y	4.69×10^{-8}
College Gardens	15176	8378	Y	0.0012

Table 5: Comparison of pre-rehab discharge and post-rehab average discharge amounts with corresponding hypothesis test results, data set adjusted to only include rain events

From the results in Table 5, it appears that for days with no rain events (as we have defined them), in the case of College Gardens there is no significant difference between the pre-rehab daily average discharge and the post-rehab average daily discharge. This does not necessarily imply that the College Gar-

Pump Station	Average Discharge		Hypothesis Test	
	2018-2019	2020-2021	Reject Null	P-Value
Briar Cliff	25856	14897	Y	9.13×10^{-23}
College Gardens	7760	7443	N	0.26

Table 6: Comparison of pre-rehab discharge and post-rehab average discharge amounts with corresponding hypothesis test results, data set adjusted to only include days with no rain events

dens repairs were ineffective, as it is the rain events we are worried about, and the results from Table 4 do indicate that post-rehab there was a significant drop in daily average discharge on days classified as rain events.

4.3 Regression Tree Analysis

[Have a description here about what a regression tree is, how MATLAB creates the regression tree, and a description of the variables (perhaps have them listed in a table]

Separating the time series data into rain-events and non-rain events gives us a more realistic evaluation of the repair work because it allows us to compare the performance of the pump stations during times of similar weather conditions. However, arbitrarily stipulating a sharp cutoff for rain events and non-rain events results in a somewhat oversimplified version of the situation. It was our goal to improve upon this by using Matlab's machine learning capabilities. Specifically, our aim was to predict the discharge amounts for the post-rehab pump stations as if they were subject to the same climate conditions as in 2018 through 2019. For this we first establish a relationship between the post-rehab climate data and the post-rehab pump discharge amounts; this is our 'model.' Next, we provide the model with the climate data from 2018 and 2019, yielding a projected discharge amount for each corresponding day in 2020 and 2021. We could then in theory perform the hypothesis tests comparing the pre-rehab discharge with the post-rehab projected discharge amounts for an accurate assessment of the repair work.

For this we used a decision tree, specifically a regression tree. A regression tree is one of a number of supervised machine learning algorithms, designed to make predictions based on available data. The advantages of this particular machine learning method are its simplicity and ease of interpretation, compared to other methods which are more of a 'black box.' The various climatic factors which were determined to affect pump discharge are called 'predictors:' month of the year, precipitation amount, ground saturation amount, high temperature, medium temperature, low temperature, relative humidity, average windspeed, and average pressure. Precipitation and ground saturation are obvious choices; the rest were selected based on their contribution to evaporation of rainwater (or preventing of evaporation). For example, high temperatures

and low humidity and low pressure would in theory accelerate evaporation, resulting in a lower ground saturation.

The term 'decision tree' comes from the general concept of how humans make a sequence of decisions. The classification tree algorithm for categorical response variables is easy to understand: For classifying, say, a species, we look at predictors such as number of legs, presence of wings, presence of fur, etc. We ask "does this sample have four legs?" Here two branches are created, one for yes and the other for no. If the answer is yes, the next "node" on the tree would ask "does the sample have wings?" Again, two branches are created, a "Yes" branch and a "No" branch. If the answer is no, the sample is determined not to be a bird, and the next node asks "Is fur present?" If the answer is yes, we have determined that the sample is a mammal, and as we traverse more nodes we narrow down the classification to the specific species. The end result resembles an upside down tree, branching as it goes down. The regression tree is a type of decision tree used when attempting to predict a quantitative outcome (the 'response' variable, pump discharge in this case). For each predictor, Matlab looks at every data point and determines the most appropriate place to branch by comparing an error metric called the Mean Squared Error, or MSE. The data point resulting in the greatest reduction of MSE is selected as the split point for that node.

Using the model created by the regression tree, we made a hypothetical projection of what the post-rehab discharge amounts would look like if the climatic conditions were exactly the same as in 2018 and 2019. We performed a hypothesis tests for Briar Cliff and College Gardens comparing the $2018-2019$ amounts with the projected $2020-2021$ amounts; the results are displayed in Table 6:

Pump Station	Average Discharge		Hypothesis Test	
	2018-2019	2020-2021	Reject Null	P-Value
Briar Cliff	32380	17510	Y	8.91×10^{-59}
College Gardens	10208	10507	N	0.7436

Table 7: Comparison of pre-rehab discharge and post-rehab average projected discharge amounts with corresponding hypothesis test results

Here we can see that the hypothesis test does not confirm that the rehabilitation was successful in the case of College Gardens. This contradicts the conclusion yielded by our previous methods; according to our regression tree algorithm, projected average daily discharge for College Gardens is actually higher post-rehab. Why would this be the case? Unfortunately, our model is not creating very accurate predictions. The accuracy of a regression tree is evaluated by a metric called the re-substitution loss (resub loss for short): the average difference between the actual values (daily discharge amounts in this case) and the predicted values. Matlab calculated the resub loss at 4185.86 gallons per day for Briar Cliff and 5887.3 gallons per day for College Gardens. These

are very high prediction errors: $\approx 23.9\%$ for Briar Cliff and $\approx 56\%$ for College Gardens. The inaccuracy of the model leads us to question the accuracy of the hypothesis test using the predicted discharge.

There are methods for increasing the accuracy of a regression tree. The first one employed was cross validation; next was 'pruning' the tree, the standard method to reduce overfitting the model to the data, followed by cross-validation of the pruned tree. The accuracy of a cross-validated tree is called the k-fold loss and is typically higher than resub loss but is a more reliable assessment of the model; none of these methods yielded any significant improvement. Table 7 summarizes the errors we obtained:

Pump Station	Full Tree		Pruned Tree	
	Resub Loss	K-fold Loss	Resub Loss	K-fold Loss
Briar Cliff	4185.86	10103.01	4185.86	12538.3
College Gardens	5887.3	6918.73	5887.3	7116.73

Table 8: Regression tree prediction error

5 Discussion and Conclusions

References