Data Challenge 3

Yichen Wang (1368222)

Introduction

The focus of this project is to achieve constant flow in the Wastewater Treatment Plant (WWTP). The lack of constant flow results in inefficient use of pumps, causing higher energy use and maintenance costs, unnecessary workload and other risks. The current strategy, which is a more ad hoc reaction does not allow for a (close to) constant flow, as shown by Aa-en-Maas. Therefore, a new strategy is required.

Not only does the lack of constant flow cause a problem, but also the rainfall prediction. Heavy rainfall and lack of control causes overflows which bring short and long-term issues for the people and the environment. Avoiding such scenario would require the development of a more constant flow and better rainfall prediction.

This brings us to the following research question:

"What advice can be given to Aa-en-Maas in order to obtain more control over a constant flow during dry weather days, taking household water and rainfall analysis into account, and what advice can be given in order to prepare better for rainfall to reduce the amount of overflows?"

Data understanding

We divide the data into two sub-sets for two models. Namely one for the dry weather model that includes only the days where no rain has occurred, in order to calculate the average household water consumption for the different seasons. The other one for wet days, which will help us build a model that predicts the change in water level based on the predicted rainfall and the accuracy of said model. From there we filter out the features we don't need for the specific models.

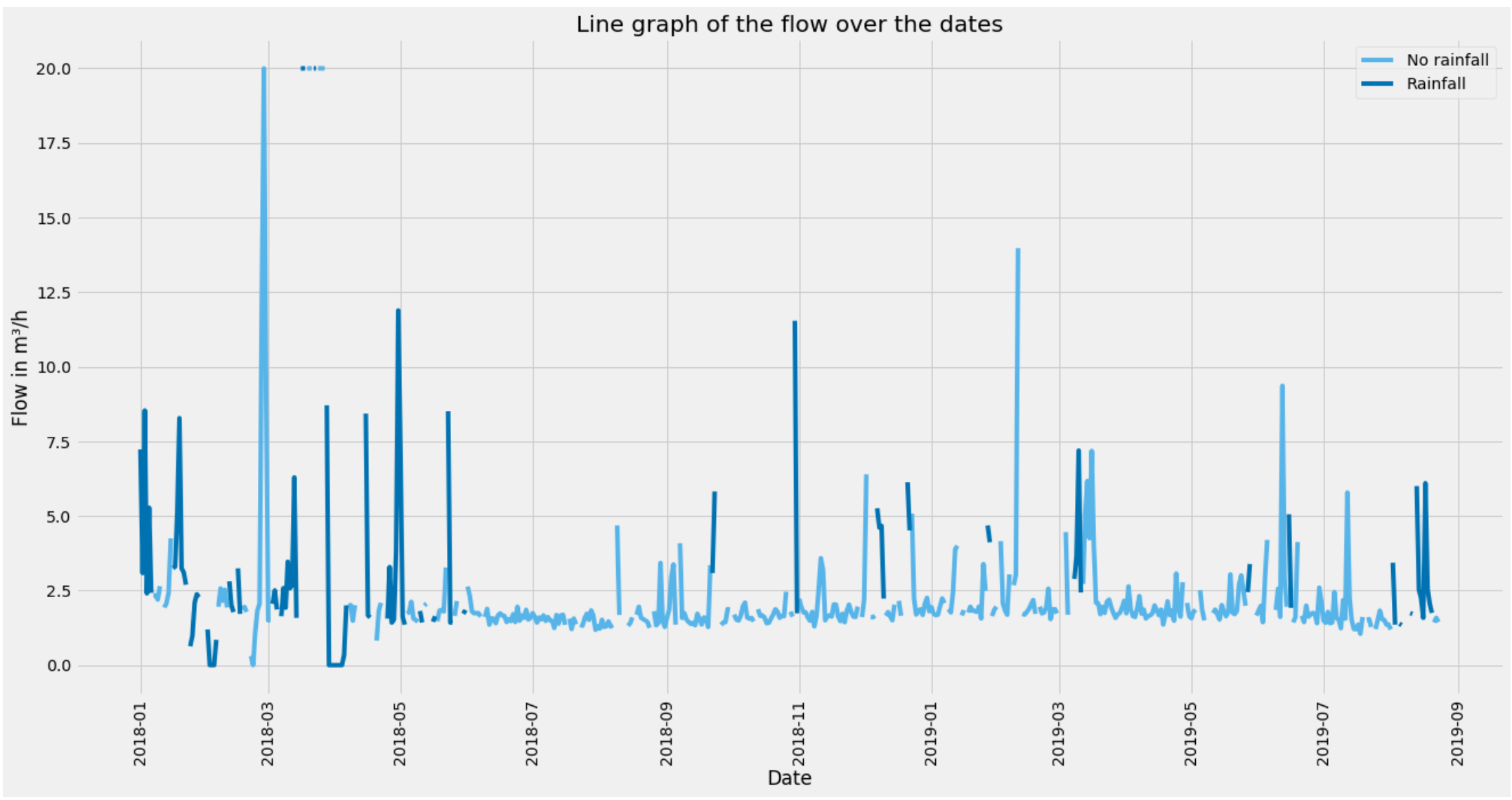


Figure 1: Flow during rainfall and dry periods

The resulting figure shows that during dry weather, the water level steadily rises on to two peaks during a day, while during a wet period the water level spikes more rapidly. A clear pattern can also be seen between the rainy periods and a spike in the flow. Our model should take this behavior into account.

The data itself has some quality problems. For example huge outliers in the data that are inexplainable or data that is missing. There is also a period where the average flow for a day is equal to zero. An explanation for this could be that the sensors had a fault in them during the measurement of the flow.

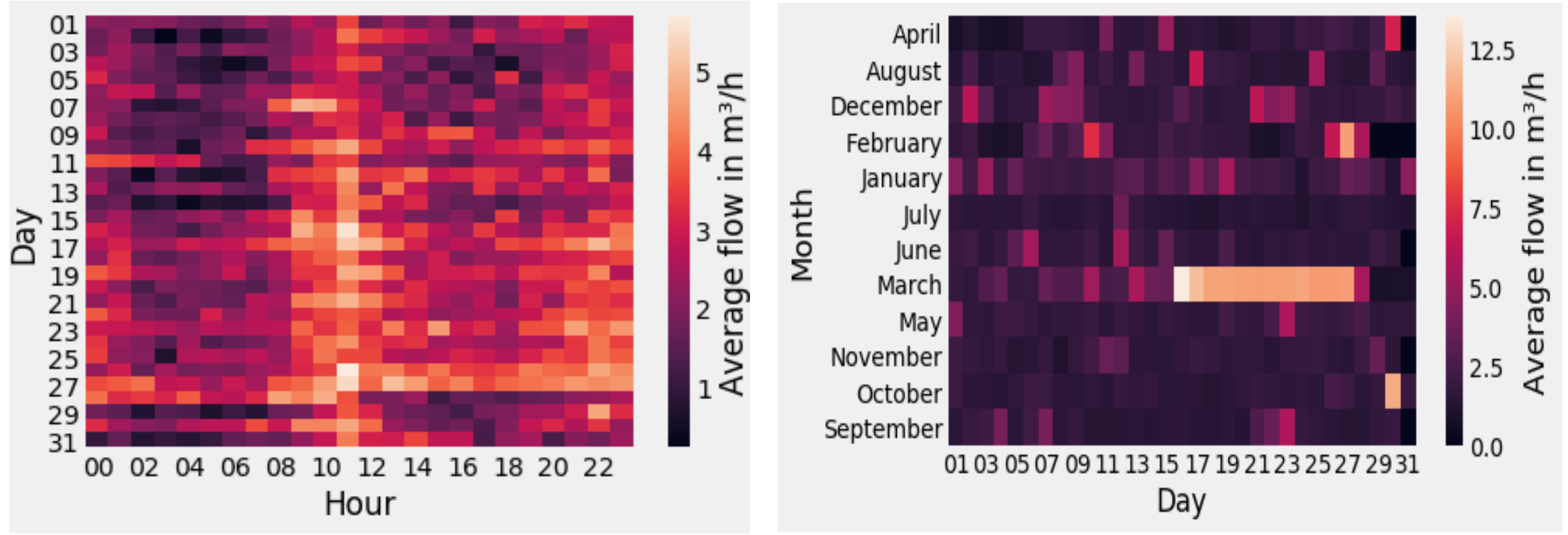


Figure 2: two heatmaps, one of the flow per day/hour and the other one for the flow per month/day

The heatmap on the left shows that the flow is at the highest from 9 am until 11 am. A reason could be that people are showering before work, restaurants using water or industries are starting to work. This results in a higher flow of the sewer system. We can also see that during the night there is barely any water usage. These patterns should be considered in our model.

The heatmap on the right confirms that there is a gap in the data from around 15th till the 27th of March. It also shows that different seasons do not have any significant effect on the flow of the sewer system.

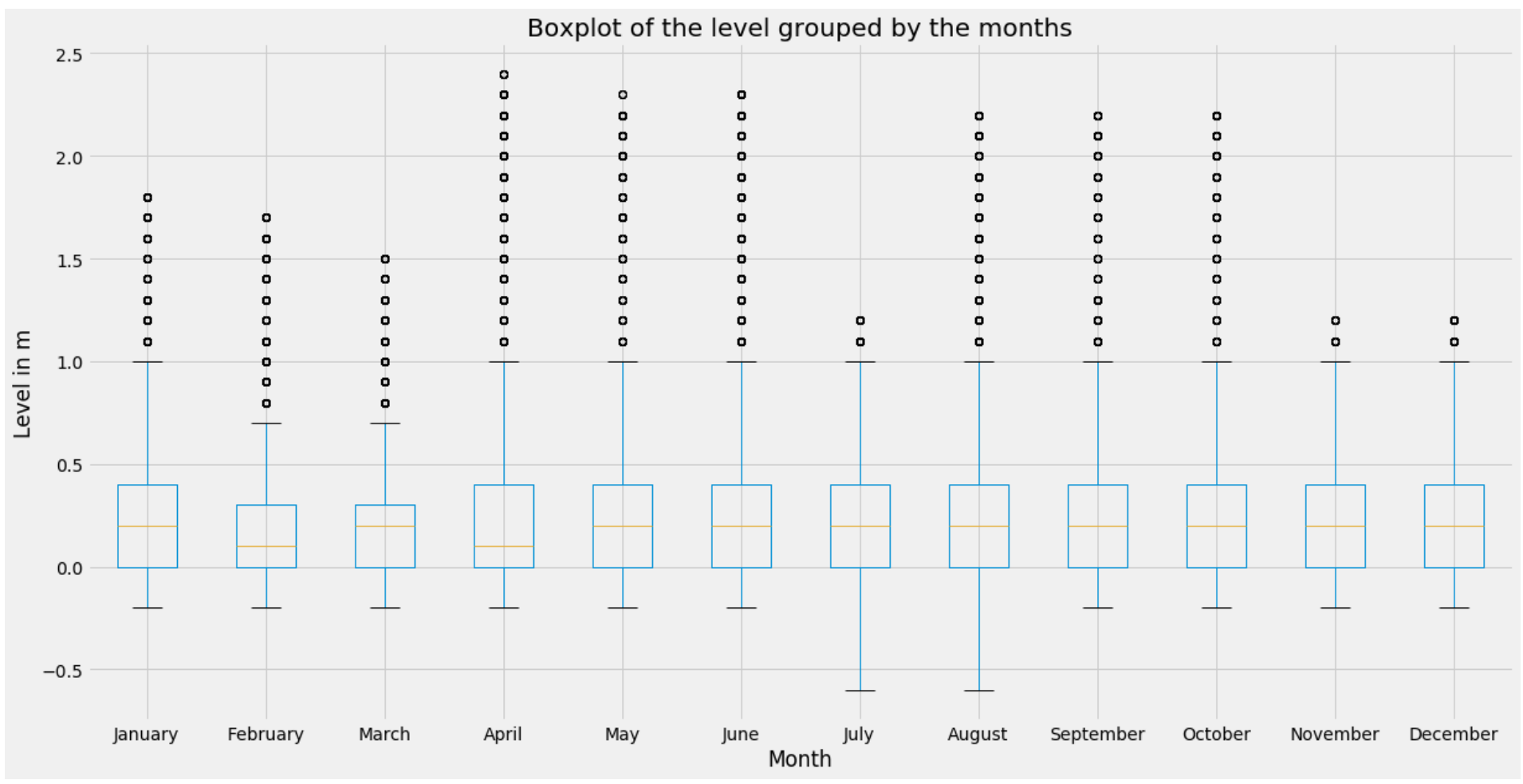


Figure 3: boxplot of the level over the months

The figure above shows the level value by month, with the months grouped on chronological order. Some months are higher than others with the largest outliers during the time between April and October. There does not seem to be a strong correlation over the year. Meaning, that there is not more problems for maintaining the level in specific months.

Figure 4 shows us that there is a higher flow during Sunday than any other day. A reason for this could be that people take longer showers or water their gardens during the weekends. Our model should take the fluctuating water usage for different days into account.

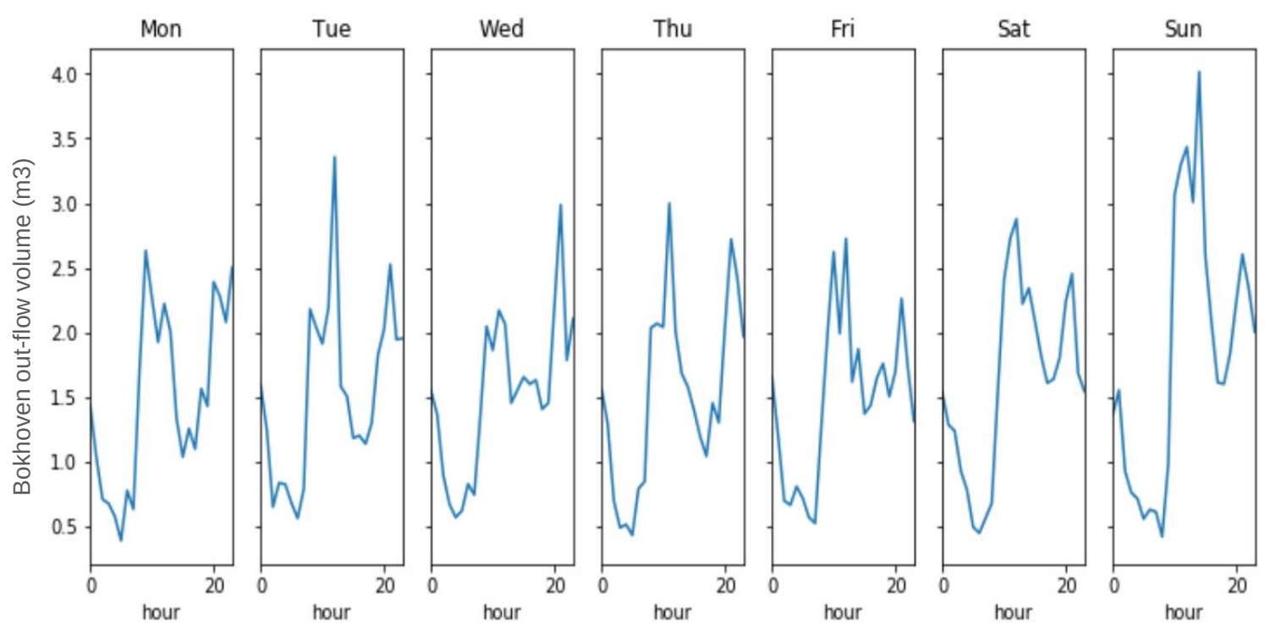


Figure 4: Flow during the days by weekday

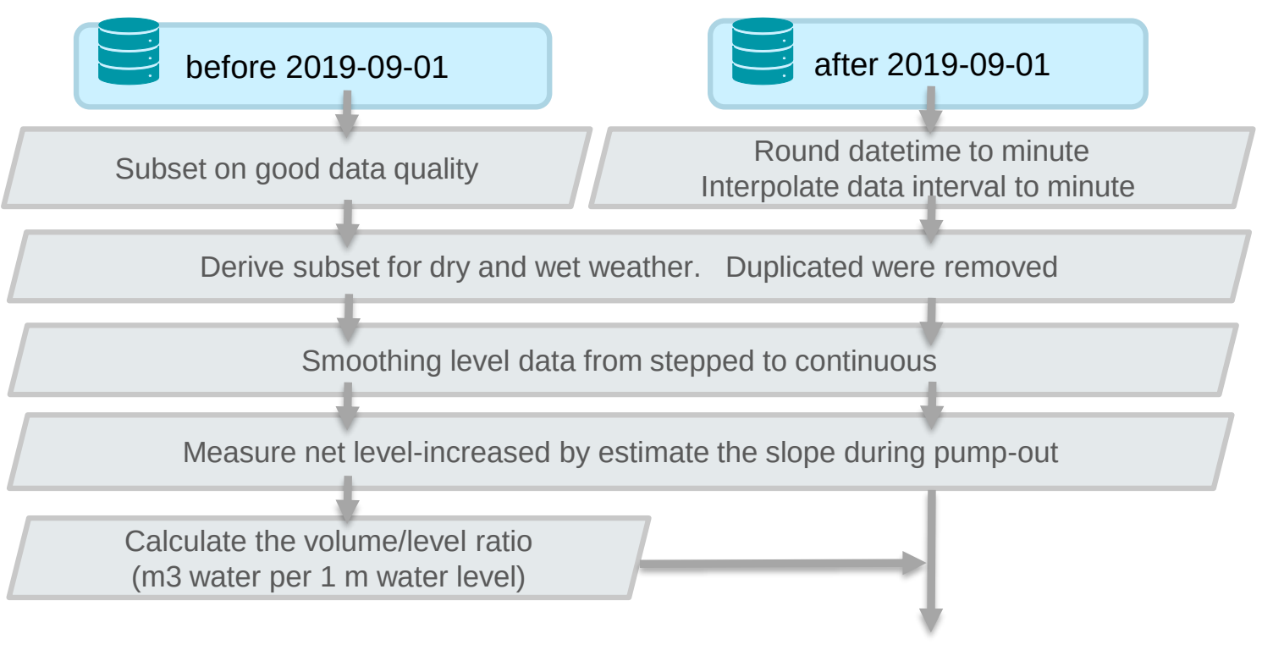
Methodology

To be able to achieve a constant in-flow to WWTP, the result of model should be able to **reduce the dispersion of total amount of out-flow from all pump stations** by operating pumps differently than previous pattern, especially for dry-day, by means of the flowing steps:

Phase 1:	**Estimate the total volume of wastewater in sewage system**
Prediction	For every individual pump station, we need: • Measure the volume of remaining wastewater from current hour • ML model to predict the volume of net in-flow wastewater for next hour
Phase 2:	**Determine a new pattern for operating pumps**
Decision	For a whole system which contains all pump stations: we need: • A mathematic model to decide each pumps if it needs switched on/off for next hour

Data pre-processing

Different pre-processing techniques have been applied to the datasets from this year and last year. The data from last year have fixed interval of time series and larger size, which is better on accurately measuring the specification, e.g.: level of switching pump, instantaneous rate of level increase, etc. Whereas the data from this year has higher precision and better time-effectiveness on later training model.


before 2019-09-01

after 2019-09-01

Subset on good data quality

Round datetime to minute
Interpolate data interval to minute

Derive subset for dry and wet weather. Duplicated were removed

Smoothing level data from stepped to continuous

Measure net level-increased by estimate the slope during pump-out

Calculate the volume/level ratio
(m3 water per 1 m water level)

Figure 5: Preprocessing of our dataset

Figure 6 shows the results of data pre-processing (the blue solid line is level data, the orange is after smoothing), and the method of volume/level ratio estimation.
As the value of water level should be continuously monotone increasing if no water pump out. Hence, the slope of level increase during the period of pump running is approximately equal to the mean value of slopes between and after, therefore, the total net level-increased could be derived accordingly to estimate the amount of water (m3) per one meter of water level increased.

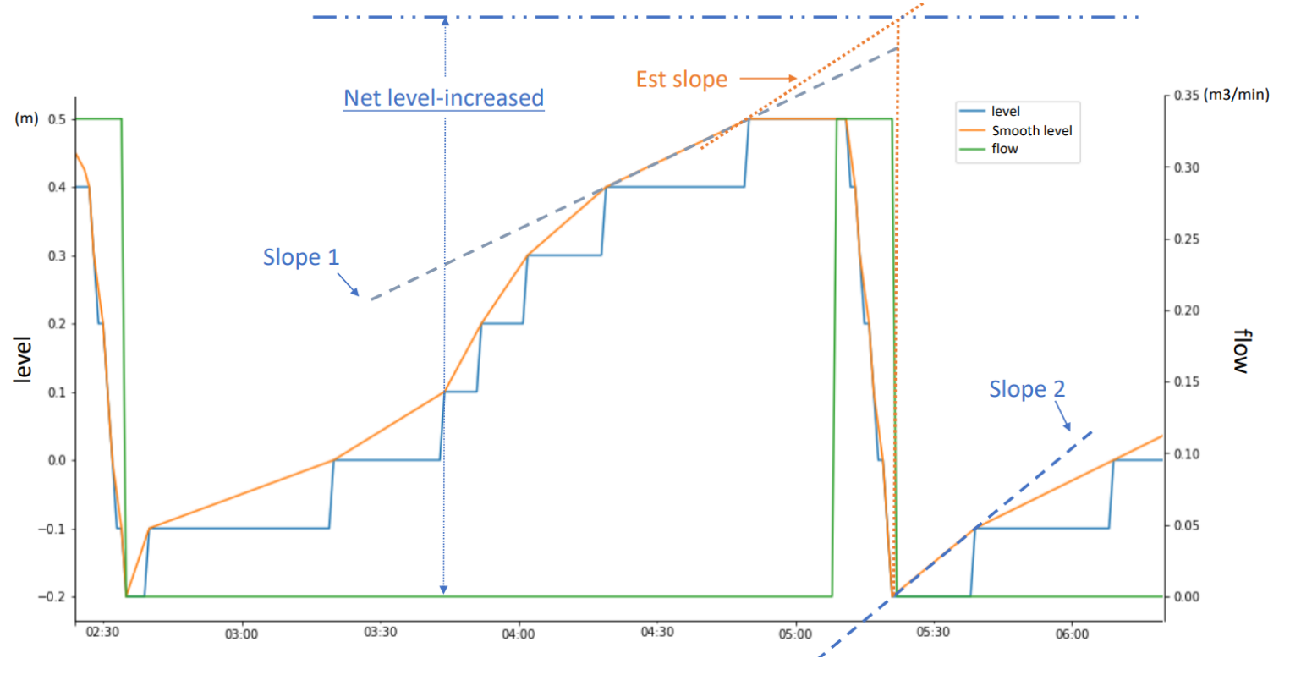


Figure 6: smoothing on level and idea of estimating

Feature Engineering

From data understanding, we have found that water flow vary with seasonality, such as the difference among months, between workday and weekend, as well as peak happens in the noon time and evening. In addition, there have researches show that water consumption is significantly higher during holiday [ref] and Covid-19 pandemic period than normal [ref a] [ref b], which refers to the uncertainties caused by human behaviour changes.
Hence, dummy features: year, month, day, weekday, hour, isHoliday, isCovidLockDown, lags of inflow, were created as candidates.
We use two steps to selecting features:

1. Using **Sequential Feature Selector** on individual dataset from three pump stations: Bokhoven, Haarsteeg, Oude Engelenseweg, accept the intersections
2. Using **independent samples t-test** to check the significance of those features (isCovidLockDown), which selected by only part of pump stations

The final selected features are: **month, day, weekday, hour, isHoliday, lag-1.**

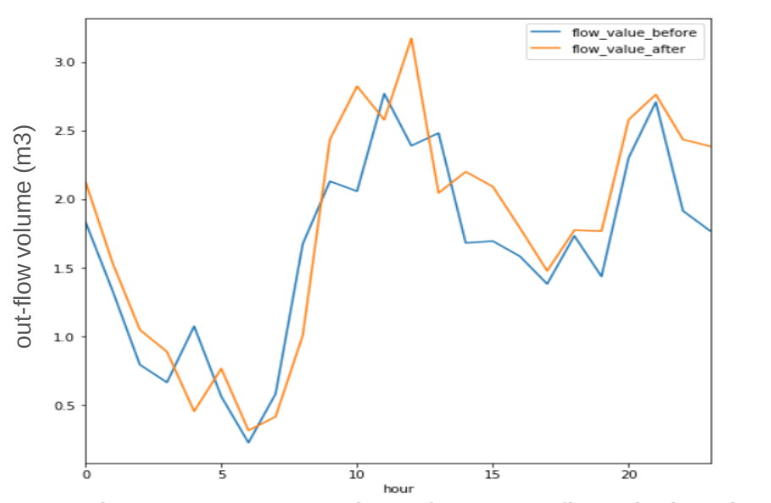


Figure 7: mean value of amount flow during the day before and after Covid-19 lockdown

Figure 7 shows the amount water flow after Covid-19 during the day is kind if higher than before. However, by two-tailed independent t-test, we have:

- T-statistic = -0.5241696031819215
- P-value = 0.6027620904711465
- Degrees of freedom = 44.49

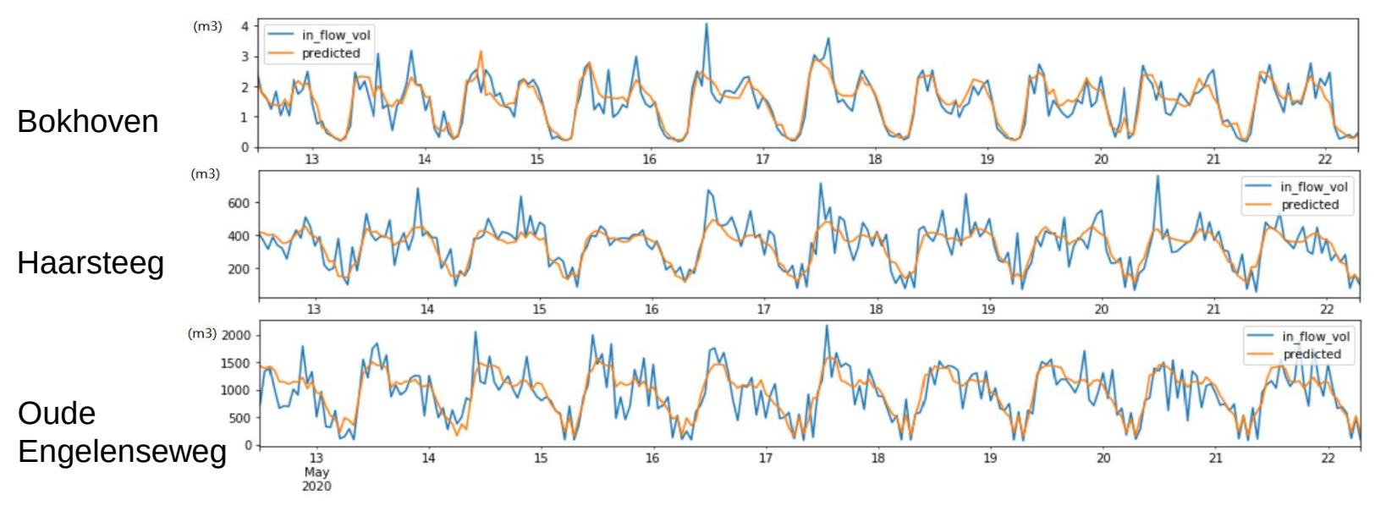
Thus, we cannot reject H0, such that there is no significant difference between two samples.

Modelling & Evaluation on In-Flow Prediction (phase1)

During dry-weather-condition, given (1) predicting in-flow water value is continuous and float, (2) those seasonality patterns reveal the strong relations between in-flow and independent variables. A regression model was decided to be applied as estimator. We took below three regression models as candidates, using GridSearchCV for tuning hyper-parameters and evaluate by MAE to get the one with best performance. Then, build predicting models separately for each pump stations.

- *RandomForestRegressor : bagging algorithm reduces variance, not sensitive to outliers*
- *XgboostRegressor: second derivative on loss function, balance variance and bias*
- *ElasticNet: less parameters, shrink variable weight by L1&L2 regularization penalty*

The plot below as example (longest continuously dry day in 2020) shows the predicted volume (orange) of in-flow water per hour by time series compare to the real value (blue) over three pump stations by the best model — fine-tuned RandomForestRegressor.



Overall, the prediction of model is pretty close to the real value in most of time, whereas slightly lacking accuracy to the extreme value during some peak hours.

Mean Absolute Error = **18.8%**
(2019/09/01 ~ 2020/08/01)

Figure 8: Predict volume of in-flow vs. Real volume per hour from 12th May 2020 to 22nd May 2020

Constant Flow Model on Operating Decision (phase 2)

Based on the results from the predicted in-flow volume, we derive the total amount of water in each sewage system and decide which pumps need to be switched on/off for the next hours. The purpose is to try avoiding all pumps running together at the same hour in order to flatten the peak and lift the pit when the sewage system has enough buffer during a dry weather day.

$flow_{mean}$: expected total flow from pumps from average of hostorical
Cap_i: capacity of water(m^3) in sewage system of i
$S_{i(t)}$: water(m^3) remaining in sewage system of i on end of the current hour t
$Vin_i(t+1)$: water(m^3) in-flow in sewage system of i in next hour, (predicted)
x_i: binary if switch-on pump i
$i \in \{all\ pump\ stations\}$

$$\begin{aligned} Min \quad & |flow_{mean} - \textstyle\sum_i [x_i \cdot (S_{i(t)} + Vin_{i(t+1)})]| \\ s.t \quad & S_{i(t+1)} + Vin_{i(t+2)} < Cap_i \\ & x_i \in 0, 1 \end{aligned}$$

Based on 0/1 Knapsack problem, we formulate a mathematical model of integer linear programming to decide if individual pump stations need to be switched-on or buffer for the next hour(t+1). It is subject to not exceeding its sewage capacity after 2 hours (t+2), to minimize the difference between the total amount of out-flow from all pumps and historical average.

Figure 9 shows the real value of the total out-flow amount from Bokhoven, Haarsteeg and Oude Engelen (blue) compared to the recommended amount by the operating decision made from our model (orange) over days with longest continuous dry period in 2020. Our model could effectively keep the total flow close to constant, by the **standard deviation** of the real flow and recommended flow. Over this period the values are, respectively, **294.52** vs **222.42**, where only 4 hours were not effectively flattened. Hence, our model could effectively reduce the dispersion of the total out-flow.

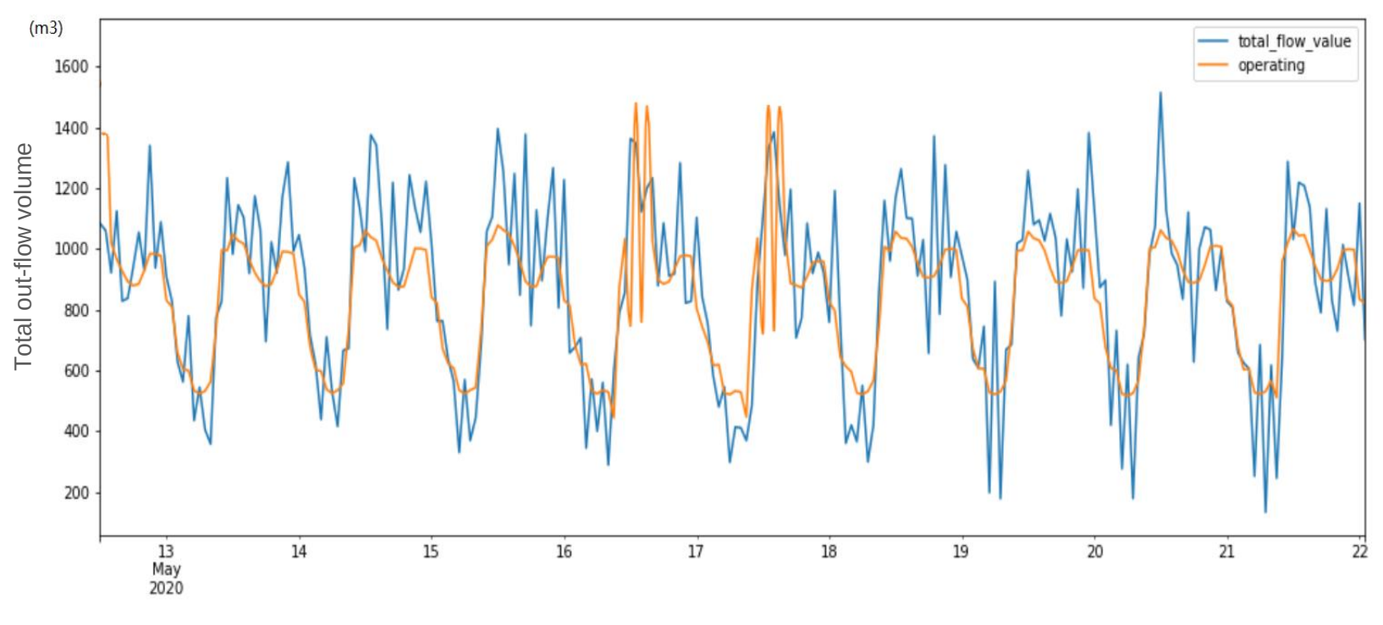


Figure 9: real amount of out-flow (blue) vs expected amount by recommended operation (orange)

Recommend buffer hours during example (total 230 hours):	
Local	**Buffer hours**
Bokhoven	*117 / 230*
Haarsteeg	*5 / 230 (5-12 12:00, 5-16 12:00, 5-16 14:00, 5-17 12:00, 5-17 14:00)*
Oude Engelenseweg	*None*

Rainfall analysis

A thorough analysis was made on the actual historic rainfall and the predicted rainfall. We plotted both values, predicted vs real and the difference (blue only, figure 10) for each of the 5 pumping station locations (Maasport and Rompert being aggregated). Afterwards we calculated the error of the predicted rainfall by taking the residuals (actual rainfall-predicted rainfall) and plotted them for each pump station. The following plots show the top 2 periods when the error of rainfall prediction was the highest.

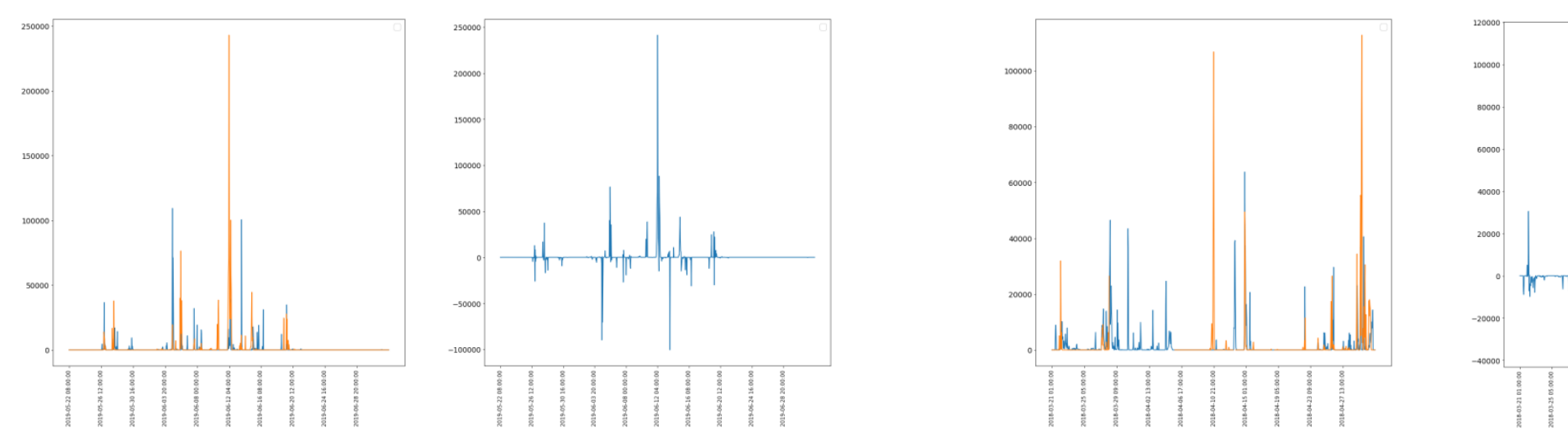
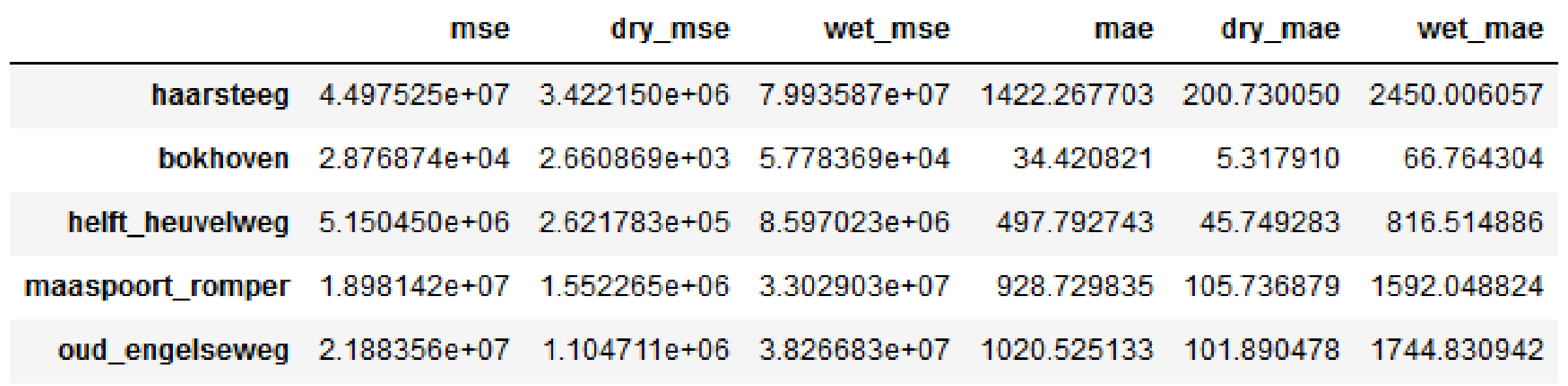


Figure 10: predicted vs real values (blue and yellow) and the difference (blue only)

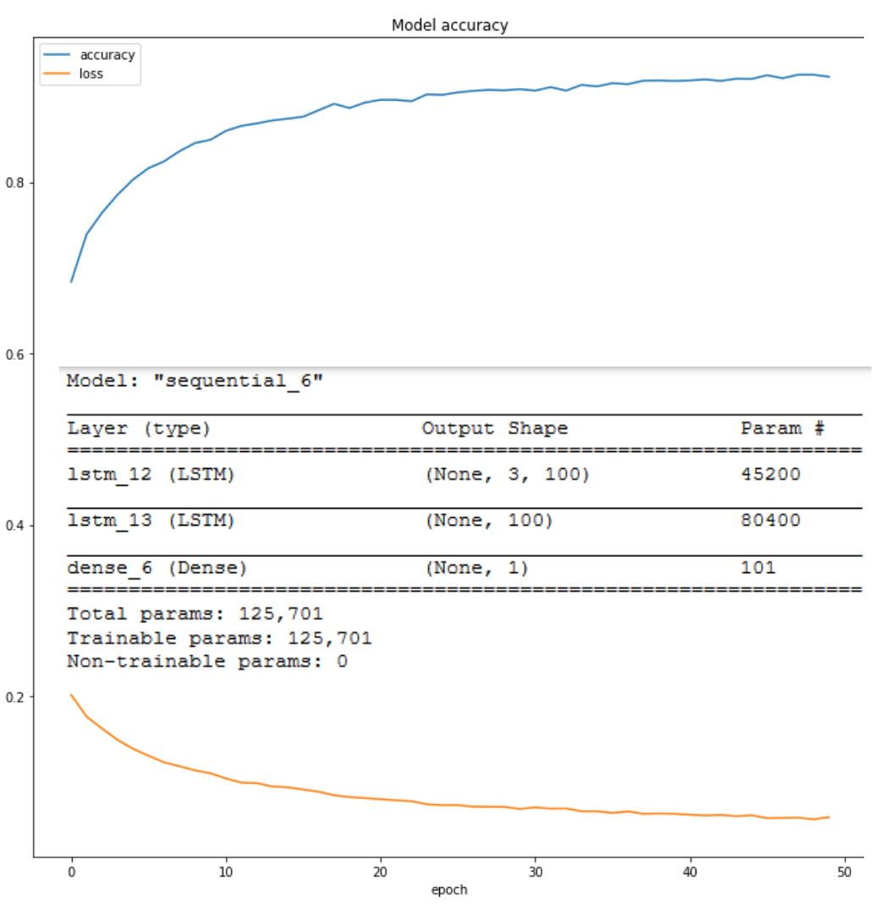
Both the Mean Square Error and the Mean Absolute Error were calculated (figure 11). MSE penalizes bigger errors versus smaller errors, while MAE does the opposite. Both MSE and MAE were calculated for all pumping stations in total, as well as for aggregations of dry/wet days for each region. From this it becomes clear that the rainfall prediction has the highest overall error for the region of Haarsteeg. In general the error is always bigger on wet days as opposed to dry days. This is yet another reason why we chose to do a dry weather model, as its uncertainty will be smaller.



	mse	dry_mse	wet_mse	mae	dry_mae	wet_mae
haarsteeg	4.497525e+07	3.422150e+06	7.993587e+07	1422.267703	200.730050	2450.006057
bokhoven	2.876874e+04	2.660869e+03	5.778369e+04	34.420821	5.317910	66.764304
helft_heuvelweg	5.150450e+06	2.621783e+05	8.597023e+06	497.792743	45.749283	816.514886
maaspoort_romper	1.898142e+07	1.552265e+06	3.302903e+07	928.729835	105.736879	1592.048824
oud_engelseweg	2.188356e+07	1.104711e+06	3.826683e+07	1020.525133	101.890478	1744.830942

Figure 11: MSE and MAE rain analysis

From this analysis the question stands if we can help improve the rainfall prediction in some way, in order to know when it's safe to use the dry weather model. For this we created a model that predicts if it's going to rain at all or not for the next 12 hours (this is a yes and no answer). The prediction is made considering the amount of rainfall for the previous 12 hours, the error of the predicted rainfall for the previous x hours and the predicted rainfall for the next 12 hours. We use an LSTM model with two very deep layers and one dense classification layer. With a loss function of MSE and an RMSprop optimizer we get an accuracy of 92%. We also created a confusion matrix (figure 12) in order to study the uncertainty of the model. We can see from the following plot that the bigger part of the misclassified values are false-negatives and although it is a small percentage this can create a problem for Aa-en-Maas, so they should use this model with the knowledge that it has an accuracy of 93% and that in those 7% an overflow can possibly happen. In general 90% of wet weather predictions are correct, 10% are false and 97% of the dry weather predictions are correct and 3% are false. This comes as no surprise as the error of predicting wet days is generally higher than dry weather predictions. One interesting insight is that 70% of the false-negatives are to data from 2018. No correlation could be made as to a specific month or day where there are abnormally more false-positives.



```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_12 (LSTM)               (None, 3, 100)            45200
_________________________________________________________________
lstm_13 (LSTM)               (None, 100)               80400
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 101
=================================================================
Total params: 125,701
Trainable params: 125,701
Non-trainable params: 0
```

Figure 12: LSTM model

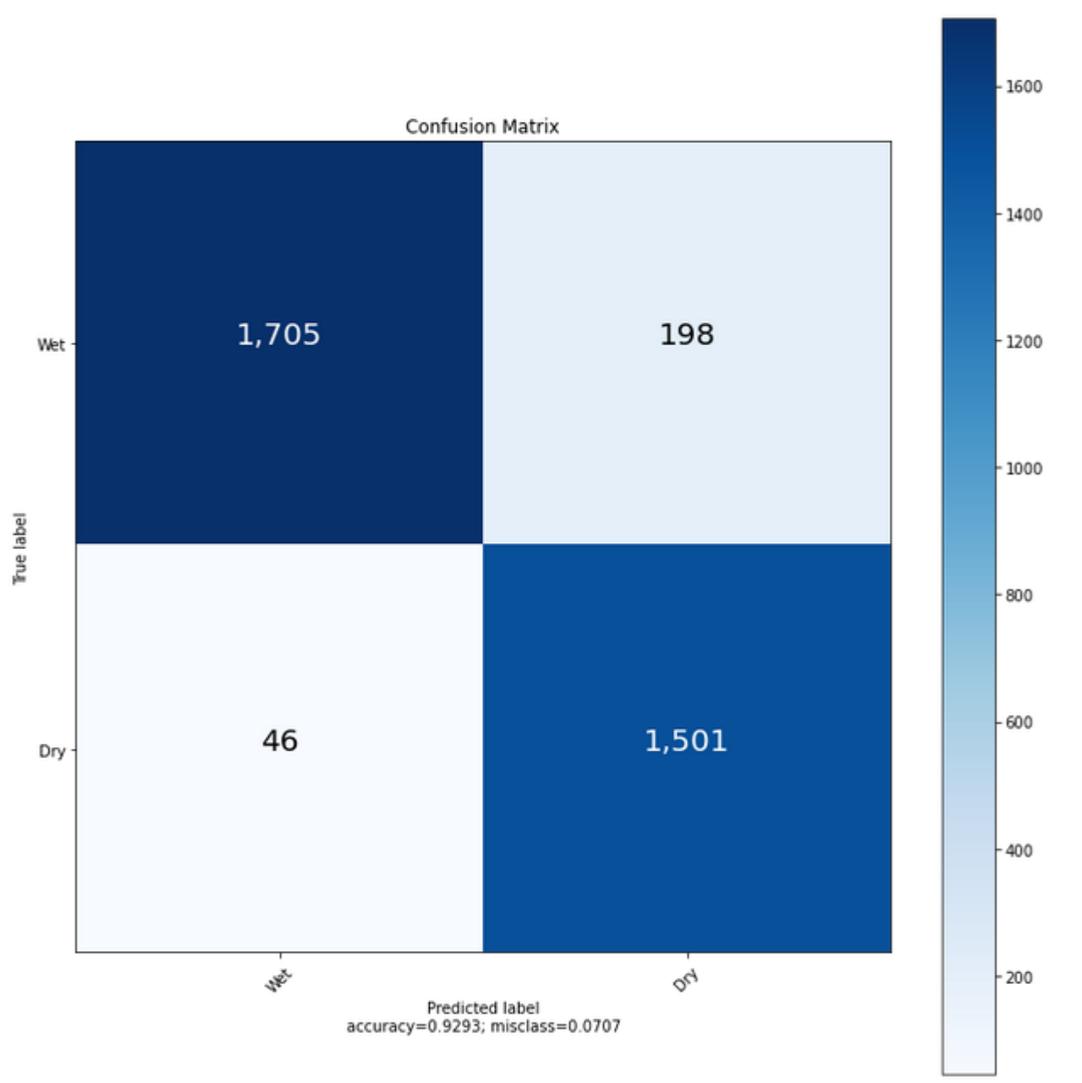


Figure 13: confusion matrix to measure uncertainty

Discussion

By the recommended operation above, the decision model mainly buffers Bokhoven half of the time for fine adjustment in order to closely approach the mean value of historical flow. Further, it occasionally buffers Haarsteeg to flatten the spike when there is an abrupt increase in the amount of water at peak hours (12:00). On the other hand, Oude Engelenseweg should never buffer due to its contribution to most of the out-flow. This pumping station has less sewage capacity compared to other pumping stations.

Concerning whether the dry weather model can be applied, more data is needed, especially the rainfall prediction. After constructing a model to predict whether it will be raining or not, adequate results were obtained. The model seems to have an accuracy of 93 percent, whereas the majority of the wrong predictions consists of false negatives. This might be a problem, but as AA&Maas has the possibility to switch back to the old strategy immediately, these mistakes can be approached by the old strategy. Meaning, that for the minority of the predictions, Plan B (the old strategy) should be followed.

Conclusion

How certain can we be about our advice? If we look at the accuracy of the rain analysis model, 93 percent appears to be correct. From the remaining 7 percent, the majority is predicting dry weather when it is going to rain. This is worse than predicting rain when there is dry weather. For these times, we suggest to switch back to the old strategy.

Another limitation to our rainfall analysis is the recency of the data. The provided data is from January 2nd, 2018 until July 31st, 2019. Therefore, to be more certain about the model and its performance, we would need more recent data.

The third limitation is that we used the rain analysis on 5 out of the 7 pumps, which makes it less reliable. However, considering the 93 percent accuracy, it has potential to be even higher.

The fourth limitation we encountered is the fact that our dry weather model is running on three of the seven pumps. We managed to show that we can accomplish a more constant flow between these three pumps, but for the other four pumps not much has changed yet. This means that there will be a partly more constant flow, but there is still a lot more potential to be reached. As our model seems to show some improvement in the flow, we are quite certain that once our model is implemented on the other pumps, the overall flow will be much more constant.

The last limitation is the performance of our dry weather model. The recommended outflow based on the inflow has errors. In the corresponding time period, the total predicted in-flow and the sewage is 198745.75 m3, whereas the total real out-flow and the sewage is 200611.68 m3. This uncertainty might be more severe during peak hours, when the use of water is increased and more pumps are involved, which might lead to higher bias of errors.

Therefore, the accuracy of the model, the data used by the rainfall prediction model, the performance of the dry weather model and the amount of pumps are still uncertainties which will be addressed in the near future.