iCASP

Integrated Catchment Solutions Programme

Living with Water Partnership Catchment Telemetry Integration

v.1

April 2021

UKRI Natural Environment Research Council

Document Title: **Living with Water Partnership Catchment Telemetry Integration**

iCASP Project: **Living with Water Partnership Catchment Telemetry Integration**

Authors: **Ehsan Kazemi, Vanessa Speight, Virginia Stovin**

Date of Issue: **DATE**

Version: **v.1**

DOI: **NUMBER**

Please cite this document as:

iCASP is funded under NERC Grant: NE/P011160/1

Non-Technical Summary

Introduction

Hull and East Riding of Yorkshire are vulnerable to flooding. The water organisations in the region, many of whom collaborate through the Living with Water Partnership (LWWP, including Hull City Council, East Riding of Yorkshire Council, Environment Agency, and Yorkshire Water), all manage drainage and related assets that interact during flood events. Each organisation currently records and stores telemetry data from their assets for their individual use, with some limited sharing of data between partners taking place but not in real-time. Believing that bringing this data together and combining it with data-driven modelling tools would create opportunities for more efficient ways of working, this project began as a collaboration between iCASP, LWWP and University of Sheffield (UoS).

The aim of the project was to integrate the data from various monitors within LWWP across Hull and East Riding of Yorkshire, and to use advanced analytical methods to better understand existing relationships in the system, with a goal of proposing a methodology for water level forecasting which can be used as an early warning tool. To realise the aim, four primary tasks were performed:

- Data collections from partners: More than 150 time-series data sets of rainfall, water level, and groundwater level were collected.
- Data cleansing, filtering, and combining: Automated systems were designed to cleanse and combine the data. This process included filtering anomalies and unwanted noise, removing high frequency spikes, smoothing (if necessary), filling the gaps, resampling, synchronising, and deseasonalising time-series.
- Exploration of relationships within data: Machine learning (ML) algorithms were applied to explore relationships between different components of the system. Many combinations of parameters (e.g., total precipitation, rainfall maximum intensity, water level, groundwater level, etc.) and locations (e.g., water level monitors at open channel streams and sewer network) were tested and major relationships were identified. Two locations were selected as case studies for predictive modelling.
- Development of predictive models to be used as early-warning tools: Based upon the results from the initial exploration of relationships, several candidate types of ML predictive models were developed, tested, and refined to select the best performing algorithm and required data inputs for an early-warning tool.

Project Approach

Given the complexity of the drainage system in the LWWP area, which includes multiple surface water features interacting with buried drainage networks, a goal of this project was to explore whether data-driven analysis could provide insights into the system performance and dynamics that are otherwise difficult to model mechanistically. Initial work to explore the

relationships in the combined data used a type of ML algorithm which is an 'unsupervised learning approach' meaning that no predefined relationships are specified within the analysis. Various combinations of parameters (total precipitation, maximum rainfall intensity, water level, groundwater level, etc) in various locations were analysed to look for trends and relationships. The findings from this data exploration task, including insights into the relationships, choice of input/output parameters, and potential time-lags between rainfall and water level in the system were used as a basis for development of predictive modelling algorithms.

The predictive models were designed based on a different type of ML known as a 'supervised learning approach'. In this approach, input and output parameters are specified in advance, thus allowing for the relationships between those parameters to be quantified (learned), and the output can be predicted for unseen (future) data. Water level a few hours in the future (in an open channel stream or a trunk sewer) was selected as the output that the model is predicting. A combination of various parameters from the preceding several hours including total precipitation, water level in the same watercourse, water level in an upstream watercourse, gradient of water level profile, and mean groundwater level used as input parameters. Sensitivity analyses were performed in order to find the best combination of input parameters, the best 'observation window' sizes (i.e., how far in the past the input parameters should be determined), and the most optimal 'prediction window' or 'forecast horizon' (i.e., how far in the future the water level can be forecasted with a reasonable accuracy).

Results

The developed model was tested for two locations in the drainage network; an open channel watercourse, Setting Dyke, and a trunk sewer in Hull West (at the monitoring location LM03). The major findings were as below.

- The model can generate forecasts for individual locations based upon historical rainfall, water level, slope of water level change, and groundwater level data.
- The farthest in the future that water level can be predicted with a good accuracy is 3 to 4 hours for Setting Dyke (open channel) and 45 to 60 minutes for LM03 (sewer).
- Combining input data from several locations, e.g. upstream open channels and groundwater level, improves the forecast horizon for the sewer network.

This project was carried out with close contact with project partners. Several meetings were held between the UoS team, LWWP members, and iCASP; extensive email communication was done for data collection; and three workshops were held with members of the project partners as well as external organisations such as Stantec and University of Hull.

The outcomes of the project are as follows:

- A better understanding of the existing telemetry network across the study area was achieved, including insights into the best locations for monitoring and gaps in coverage.

- The results of the data exploration provided clarity on the variability of data in quality and physical parameters.
- The data-driven approach was able to identify important relationships between network elements such as that the water level in both open channels and sewer network show stronger linkage with total precipitation rather than maximum rainfall intensity; and that the water level in the sewer network responds much faster to rainfall.
- Overall, application of data analytics like those demonstrated will make better use of current monitoring systems and provide evidence to support future investments.
- It is possible to improve flood resilience in the area through application of an early warning tool.
- The value of combining and sharing data among the different LWWP partners was strongly demonstrated along with the value of data-driven methods to help understand the behaviour of complex systems.

Applications and recommendations

The output of the project is a predictive modelling approach with developed models coded in MATLAB. Since the predictive model uses past values of data to predict water level in the future, it can thus be used as an early warning tool. The application of these models in real-time water level forecasting requires training of the model using existing historical data (updated in real-time), and then input of values for rainfall, water level, and groundwater level to predict water level in a watercourse a few hours in the future. For replication of the approach beyond the two tested locations evaluated in this project, individual site models will need to be trained and tested based on their local system data, as the relationships discovered for LWWP in this project will not be universally applicable.

Besides, for applications beyond LWWP, to be investigated in the future studies, the following recommendations are proposed:

- Data collection in a more systematic way, ideally in unified measuring systems.
- Thorough sensitivity analysis on the hyperparameters of the ML algorithms.
- Incorporating flood risk into the model to estimate risk of exceedance of water level triggers.

By performing these recommendations, the developed system can then be used for more informed decision makings. This can help more proactive interventions at operational levels (e.g., identification of immediate risks to the system) and strategic levels (e.g., assessment of large areas for overall risk of water level exceedance above defined thresholds under typical high-rainfall conditions).

Contents

1. Introduction

According to a survey[1] conducted by the University of Hull, Hull and East Riding of Yorkshire are very vulnerable to flooding and were severely affected by major flood events in 2007 and 2013. In June 2007, surface water flooding in Hull damaged approximately 8800 residential properties, 1300 businesses and 91 out of 99 schools. In December 2013, a storm surge flooded over 400 properties in Hull and East Yorkshire. Therefore, Hull City Council (HCC), Yorkshire Water (YW), the Environment Agency (EA) and East Riding of Yorkshire Council (ERYC) have formed the Living with Water (LWW) partnership to work together to reduce vulnerability to flooding and increase resilience in Hull through infrastructural projects and at a community level.

Currently, each of the LWW partnership organisations (HCC, ERYC, EA and YW) record and hold telemetry data from their assets in Hull and the surrounding region, something that is replicated throughout the UK. Each organisation uses this data for their own purposes and responds on an individual basis. The LWW partnership believes that by better utilising this data there are opportunities for more collaborative and efficient ways of working. Therefore, in this project, we have brought together data sets from different organisations, combined them, and used advanced Machine Learning (ML) methods to highlight relationships within the data, and develop predictive models to be used as 'early warning' tools using data that is already available. These models will predict water level in watercourses using historical data of rainfall, water level and groundwater level. Additional uses of the analysis will highlight opportunities for further monitoring and data collection as well as provision of public information.

2. Aims and Objectives

As mentioned above, the aim of the project is to integrate the data from the various monitors within LWW partnership and to use advanced analytical methods to better understand existing relationships among the parameters, with a goal of improving operational performance with this knowledge. The tasks performed in the project are as below.

1- data collection from partners
2- data cleansing, filtering and combining
3- exploration of relationships within data

[1] Hull Household Flooding Survey 2018, Energy and Environment Institute, University of Hull, (online) available at www.hull.ac.uk/editor-assets/docs/hull-household-flooding-survey.pdf, accessed on 15 April 2021.

4- development of predictive models to be used as early-warning tools

In the following sections, some details of analysis, results and findings corresponding to the objectives of the project are presented in Sections 3 to 7; then in Section 8, some recommendation regarding how to implement the model and replicate it for other areas will be provided; and finally, in Section 9, the overall summary and conclusions are presented with recommendations for future studies.

3. Data Collections from Partners

LWW partnership organisations provided more than 150 time-series of rainfall, water level and groundwater level data from monitoring stations across Hull and East Riding of Yorkshire. The water level data includes water level in open channel watercourses all around the East Riding of Yorkshire as well as trunk sewers in the city of Hull. A full list of data provided by project partners is presented in Appendix A, and a subset of data used in the analysis presented in this report is shown in Table 1. Note that the whole dataset presented in Appendix A were looked into, most of them were cleansed, but not all of them were used in the analysis due to the issues discussed below. Also note that the list presented in Table 1 shows only the ones employed for the analysis reported here, while a part of analysis is not presented in this report for the sake of brevity.

Table 1: A list of a subset of data which is used in the analysis presented in this report. For a full list of data collected from project partners see Appendix A.

ID	Type	Location/Tag	Easting	Northing	Organisation
TS1	Rainfall	Cottingham	504791	434188	EA
TS2	Open channel water level	Hessle Western Drain	503295	427583	EA
TS3	Open channel water level	Setting Dyke (National Ave)	506692	430535	EA
TS4	Groundwater level	Cottingham Willerby Hill	502281	431806	EA
TS5	Groundwater level	Cottingham North House	505000	435000	EA
TS6	Open channel water level	Atwick Village Drain	518940	450847	ERYC
TS7	Open channel water level	Bilton	517221	432531	ERYC
TS8	Open channel water level	Plaxton Bridge	506611	436548	ERYC
TS9	Open channel water level	Skirlaugh	514217	439770	ERYC
TS10	Open channel water level	Astral Close Screen	503670	427320	ERYC
TS11	Sewer water level	Swanland	500433	427678	YW
TS12	Sewer water level	Dawson House	508363	431688	YW
TS13	Sewer water level	Hull West (LM02)	506563	431909	YW
TS14	Sewer water level	Hull West (LM03)	509415	429239	YW
TS15	Sewer water level	Hull East (LM04)	510791	429472	YW
TS16	Sewer water level	Hull East (LM05)	512969	430919	YW

The data is not uniform in terms of duration and frequency of measurement, continuity, etc. For example, the data from most of the YW rainfall gauges has a frequency of one sample per day which make them insufficient for the analysis; while the data from an EA rainfall gauge in Cottingham (TS1 in Table 1) has a time interval of 15 minutes and a measurement duration of about 35 years which makes it the most useful rainfall data for our analysis. The EA groundwater level data also lacks sufficient sampling rate in most of the cases. Besides, in many cases, there are issues such as that data resolution is varying over time, there are large gaps in the data, the duration of data is not sufficient, the amount of unwanted noise and high frequency spikes is large so that the data cannot be cleansed, and so on.

Figure 1 presents the monitoring location and type of data used in this report (see Table 1); and Figure 2 shows duration (length of time over which data is measured) and resolution (average number of samples per day) of this data. For instance, rainfall at Cottingham (TS1 in Table 1) has large duration and good resolution, groundwater level at Cottingham North House (TS5) has a relatively large duration but its resolution is poor, and water level at Astral Close Screen (TS10) is poor in terms of both duration and resolution.

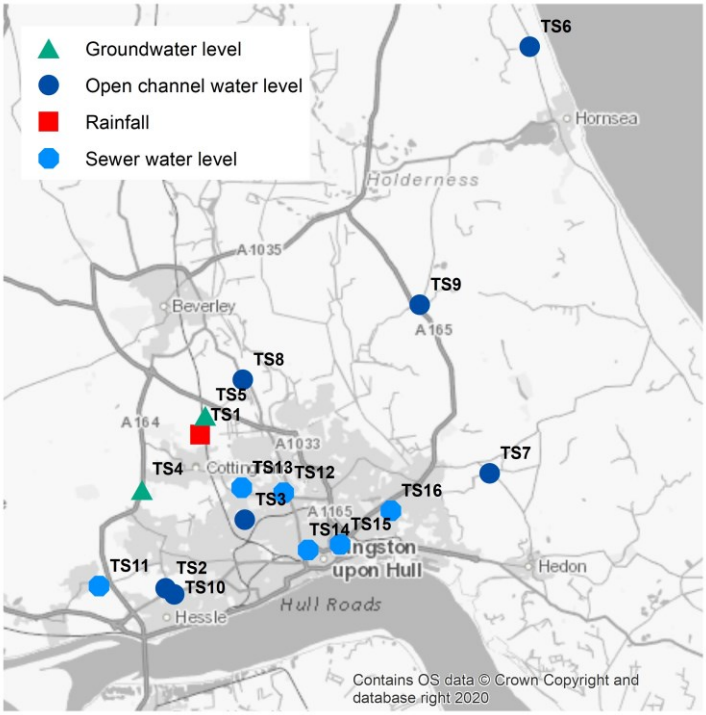

Figure 1: Location of monitoring stations of the data used in the analysis presented in this report (see Table 1).

According to the issues with the raw data discussed above, the data needs filtering and cleansing before being used in the ML analysis. In addition, the data from different sources need to be organized and combined into a single dataset that can be employed by the ML

models for exploring relationships between different elements of the system. The process that is used for filtering, cleansing and combining data is demonstrated in the next section.

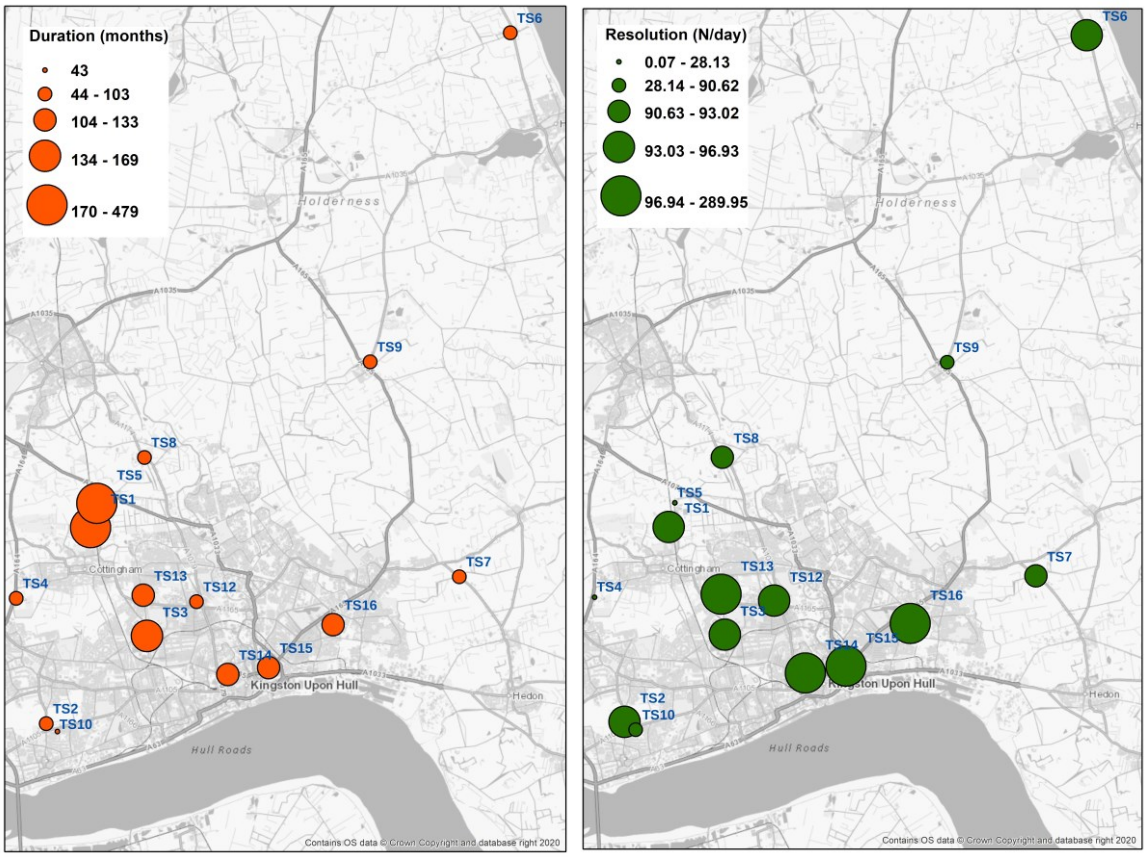

Figure 2: duration (left) and resolution (right) of the data used in the analysis presented in this report (see Table 1). Duration denotes the length of time over which data is measured at a monitoring station in months, and resolution denotes the average number of measured values per day.

4. Data Cleansing and Combining

Cleansing data includes all or some of the following processes on the time-series.

a. Removing spikes and unwanted noise
b. Filling gaps
c. Smoothing
d. Removing seasonality
e. Resampling

An automated system is designed in MATLAB 2019b to perform all these processes automatically since the dataset is very large and manual removal of anomalies, filtering/smoothing, etc. is not practical. High frequency spikes (which could be due to Instrumental errors) are considered as outliers and are filtered out; and if there is unwanted

noise in the time-series, it is filtered or smoothed using ‘medfilt1’ and ‘smooth’ functions in MATLAB R2019b. Figure 3 shows an example of removing spikes from and filtering a water level time-series (TS7 in Table 1), where blue line shows the raw data and red line shows the filtered time-series. For some of the analysis (presented in Section 6) seasonality of data is also removed. If there are unwanted seasonal components in the time-series, they are firstly modelled using curve-fitting or smoothing methods. The time-series is split into a number of segments and then at each segment polynomial curve fitting or smoothing functions are applied and then by combining the segments the seasonality is modelled and then removed. Figure 4 shows an example of this process for a water level time-series (TS6 in Table 1), where blue, green, and red lines show data (raw or filtered), seasonal component, and deseasonalised data, respectively. The output (red line) in this case represents changes in water level from the base value (green line) to water level (blue line).

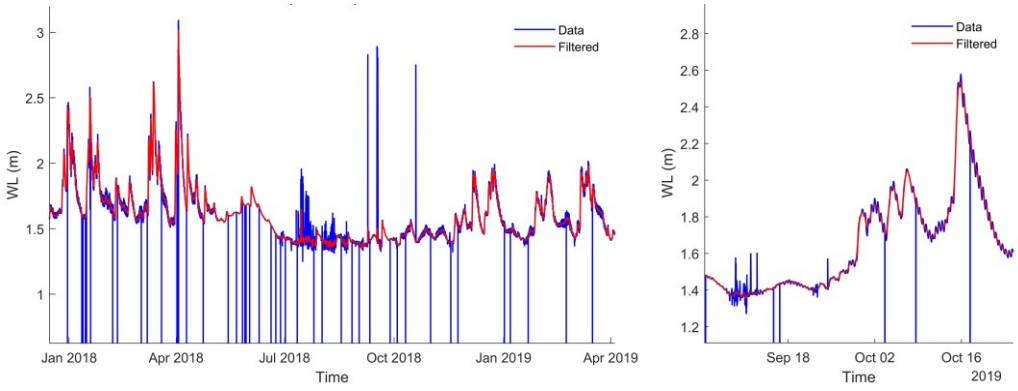

Figure 3: An example of removing high frequency spikes from an open channel water level time-series (TS7 in table 1). Blue and red lines show raw and filtered data, respectively.

Another issue in the data is that due to being collected from different sources and data types (rainfall, water level, groundwater level, etc.), the frequency or sampling rate of data is not the same for all time-series. Therefore, they first need to be resampled at a certain frequency to be combined before being used in the analysis. For the predictive models, the higher the data frequency is (i.e., the smaller the time interval is), the higher the resolution/accuracy of the predictions can be achieved. However, it is not possible to resample the data at a very large frequency as we wish, in other words, we are limited to a certain range. This is because on the one hand, the smallest time interval of the existing rainfall data is 15 minutes, and on the other hand, there are time-series with very large time intervals which cannot be resampled to very high frequencies without losing information. For instance, the data from many rainfall gauges in the region has frequencies of around one sample per day, and if we want to use them for prediction of water levels with frequencies of around, for example, one sample per hour, we will then need to interpolate rainfall values to fill the one-day gaps between each two

successive points in the time-series. But this is not ideal for such data since rainfall events are often in the order of several hours and by performing such interpolation, we will still lose many of the events. However, for some other time-series, such as groundwater level data, interpolation between two values with a time interval of even a few days is still OK since groundwater level often changes very slowly, and thus we will probably not lose any meaningful variations between the two successive points in the data.

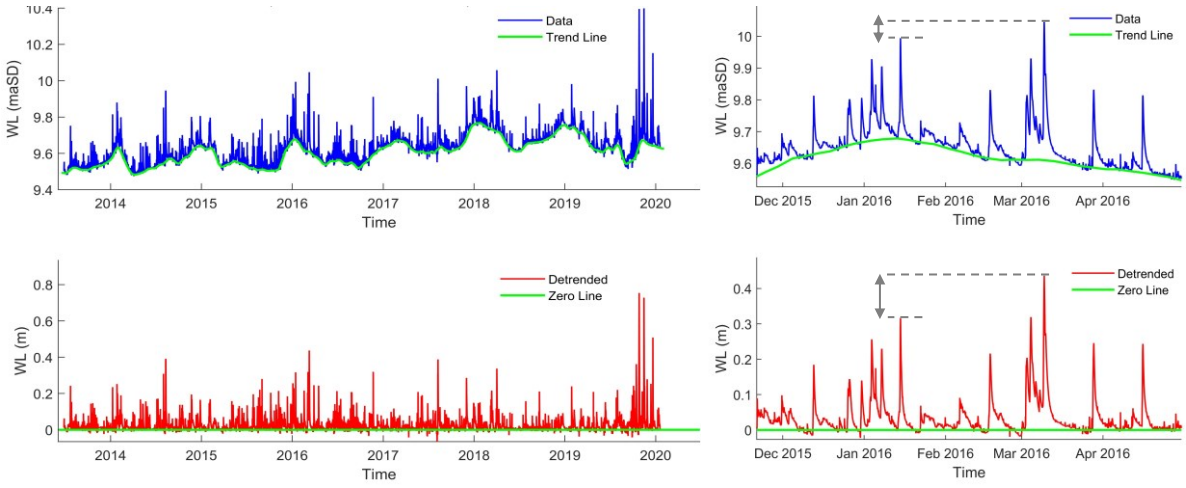

Figure 4: An example of removing seasonality from a water level time-series (TS6 in table 1). Blue, green, and red lines show data (raw or filtered), seasonal component, and deseasonalised data, respectively.

In the present analysis, rainfall data from a gauge in Cottingham (TS1 in Table 1) is used as the only rainfall input since it is measured at a frequency of one measurement per 15 minutes and has been measured for a period of 35 years since 1985. Therefore, this time-series is set as the reference time-series and all the other data (water levels, groundwater levels, etc.) are synchronised to it. That means frequency of the input data fed into the ML models is 15 minutes. Meanwhile, water level data with low frequencies (lower than one measurement every one hour, which makes them not suitable for resampling by interpolation) are not employed in the analysis.

5. Machine Learning Introduction

The present data is sparse in space and time, and data linkages across functions are lacking. Therefore, data-driven machine learning, which map inputs to outputs without attempting to accurately model underlying processes, seems as one of the most suitable choices for this study. Thanks to the availability of historical data recorded by the LWWP organisations, and due to the lack of detailed knowledge on the complex physical processes in the system, data-driven models are developed in this study for exploration of such relationships and trends in

the data and then predictions of water level in open channel watercourses and trunk sewers in Hull.

Unlike the traditional statistical methods that work based on a priori assumptions about data (such as linear and repeatable trends) and that are therefore suitable only for small datasets with straightforward and stable relationships, ML, by enabling the system to learn from data, yields better performance when data is large and relationships are complex and/or nonlinear. It can identify dominant mechanisms and empirical relationships in large datasets by mapping inputs to outputs without attempting to replicate assumed underlying processes, a property which has made it a useful method for various engineering applications.

ML techniques can be grouped into supervised and unsupervised approaches. Supervised ML approaches such as regression- and classification-based Artificial Neural Networks (ANN) are employed when parameters in data are divided into input and output variables and an algorithm is used to learn the mapping function from the input to the output. This is the process used in Section 7 to develop predictive models. We know that our output parameter is water level (preferably in the sewer network), but we do not know what combination of input parameters (for example, rainfall in location A, water level in open channel B, etc.) should be used to achieve the best performance, unless we do 'trial and error'. However, such trial and error with supervised ML models is quite difficult and time-consuming. Therefore, our alternative is firstly to explore relationships with an unsupervised ML technique, as they are faster and less complicated and useful for when relationships between parameters are poorly understood and prior knowledge about data (input/output parameters) is unavailable.

Two common unsupervised ML techniques are Self-Organising Map (SOM) and Principal Component Analysis (PCA), but the latter is unable to deal with missing values and nonlinear relationships between parameters, while SOM can easily handle both issues (Speight et al., 2019). SOM is a type of ANN which is suitable specifically for visualizing relationships within large datasets, especially in the presence of high-dimensionality, by producing a low-dimensional (usually two-dimensional) discrete representation of the input space (Kohonen et al., 1996).

Therefore, the following steps are undertaken to analyse the data of drainage system in Hull.

SOM is firstly applied in Section 6 to explore relationships and correlations between different parameters such as rainfall, water level and groundwater level, at various locations. The identified relationships will then create the basis for the predictive modelling in Section 7, where two supervised ML methods, namely Feedforward Neural Network and Random Forests are employed. Random Forests is applied for both classification and regression.

Therefore, three predictive models are developed for generating quantitative forecasts of water level from historical data of rainfall, water level, and groundwater level. The Feedforward model is named 'Regression FF', and the classification- and regression-based Random Forests models are called 'Classification RF' and 'Regression RF' throughout this report. Figure 5 summarises the methods applied in this study.

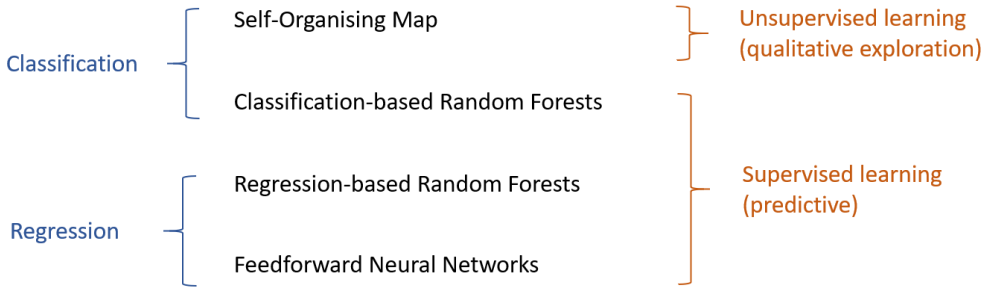

Figure 5: ML methods used for qualitative exploration of relationships within data and quantitative prediction of water level.

6. Exploration of Relationships

6.1 Introduction to SOM

By using a 'competitive learning' approach, SOM produces a nonlinear mapping from an m-dimensional space of attributes to a two-dimensional lattice of cells or neurons (Kind and Brunner, 2013). For the sake of brevity, an example of how to read SOM lattices is presented in this section and the readers are referred to the studies listed in the Reference section for further study about the technical details of the method.

As an example, Figure 6 shows the results of training of a SOM model using four years of data of rainfall in Cottingham (TS1) and water level in LM03 in the Hull West (TS14). At each 15 minute interval, total precipitation in the last 6 hours and magnitude of water level at 45 minutes in the future are calculated and fed into the model as input parameters. The relationship between these two parameters is then presented by SOMs on two-dimensional lattices. On the lattices, each cell (neuron) represents a group of observations; the spatial location of a cell corresponds to a particular domain or feature drawn from the input space; colours show the value of the variables (red: high, blue: low); and each cell in the same position on different lattices corresponds to the same group of observations/samples. For this example, SOMs show that there is a strong linkage between total precipitation in Cottingham and water level in LM03 as water level is high/low in the same positions on the maps where rainfall is high/low (high value regions are shown by dashed lines). Note that it does not matter where on the

maps the clusters form, but the important point is that the neurons (and patterns) at the same position on different maps correspond to the same group of samples in the data.

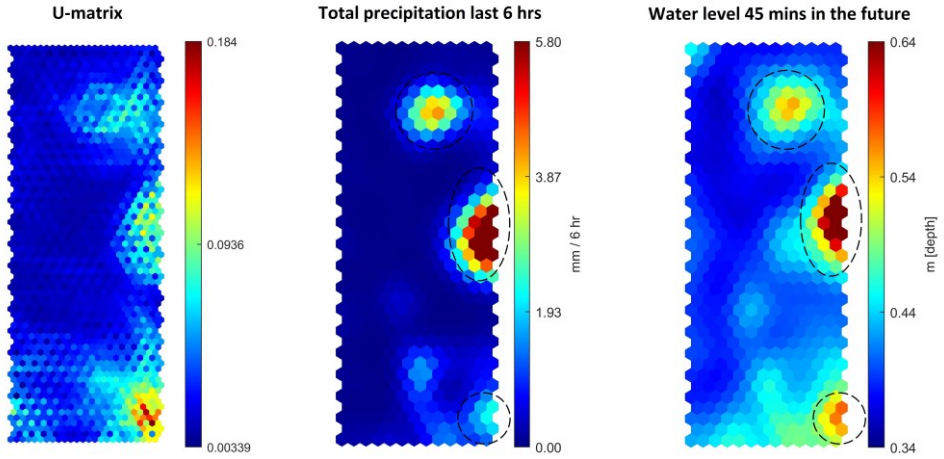

Figure 6: A simple example of SOMs representing relationship between rainfall in the last 6 hours at Cottingham TS1 (middle) and water level at 45 minutes in the future in trunk sewer LM03 TS14 (right). The map on the left shows the U-matrix.

In addition to the maps of parameters, SOM provides an additional lattice called unified distance matrix, or 'U-matrix' (see Figure 6, left). A U-matrix represents the 'distance' between clusters and determines the strength of them.

In the above example, the number of parameters was just two. The strength of the SOM method is in uncovering relationships in a data with large number of parameters and measurements and potentially several nonlinear and complex relationships.

6.2 SOM analysis of the dataset

It is easy and fast to test various combinations of data using SOMs. Many combinations of rainfall, water level and groundwater level data were tested, and it was tried to understand whether and how water level in open channel streams and sewer network are correlated to other data. For this purpose, several tests were designed in order to examine relationships in raw data, filtered data, or deseasonalised data; based on storm events or data points; for short periods of time such as a few months, or for longer periods such as 10 years; using time-series synchronised in the same time or with time lags between different parameters or locations; using combinations with small numbers of parameters/locations, or employing many parameters/locations together; and so on. Below, only a few of these analyses are presented followed by the most important findings in Section 6.3 which are then used to design the predictive models in Section 7.

One of the tests was exploring relationships between rainfall in Cottingham and maximum water level in the open channel watercourses and the sewer network. This test was based on storm events. Firstly, rainfall events were detected, and two rainfall parameters were calculated for each event, i) total precipitation (mm) which is sum of rainfall values during the event, and ii) maximum rainfall intensity (mm/15mins) which is rainfall peak in the event. Then, maximum value of water level in the next 24 hours after rainfall event both in the open channels and the sewer network was calculated. Figure 7 shows the process of calculation of input parameters. Note that, in this test, water level was deseasonalised as discussed in Section 4; therefore, 'water level' in this section means water level above baseline rather than water level itself.

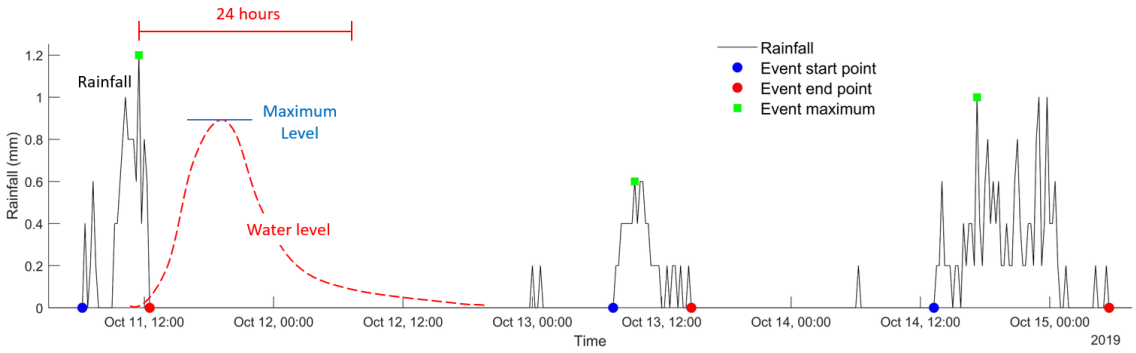

Figure 7: Design of event-based SOMs. Examples of rainfall events detected by the automated rainfall event detection model (black), and a schematic water level rise in the next 24 hours after a rainfall event (red).

For rainfall event detection, an automated system was developed which searches over the time-series, finds events with total precipitation or maximum intensity above a threshold, and then separates the events based on their distance in time, i.e. calculates the dry period between two successive events, and if it is below a threshold, they are combined into a single event, and if it is above the threshold, then they are kept as two separate events (threshold is set to 6 hours in the present analysis).

Figure 8 shows one of the SOMs for the test. Lattices A and B represent rainfall parameters, which are total precipitation and maximum intensity, respectively, in Cottingham (TS1); C to G are the maps for five of the open channel watercourses (TS3, TS7, TS8, TS9 and TS2); and H to M are maps of water levels in the sewer network (TS12, TS11, TS13, TS14, TS15 and TS16). Looking at Lattice A, a cluster of moderate total precipitation (cyan area) forms in the shape of a triangle (marked by dashed line), which is also present in the rainfall intensity (latices B); while the high rainfall clusters forms in different positions on the maps, i.e. bottom left corner on Lattice A and bottom and right sides on Lattice B. Looking at the water level maps, some of the open channels show correlation with rainfall, especially at high total

precipitations, such as Western Hessle (Lattice G) and Setting Dyke National Ave (Lattice C); but they do not show strong linkage with rainfall parameters at moderate values. This could be due to that it probably takes longer for open channels to return to their normal level after a storm if compared with the sewer network presented in Lattices H to M. Most of the sewer network levels show correlation with rainfall at moderate values (triangular shape), too. This can be seen, for example, in LM03 and LM04 lattices, where the high water level values not only links to high total precipitation in the bottom left corner of Lattice A, but also with moderate rainfall values in the triangular shape. These relationships suggest that response of water level in the sewer network to the rainfall is fast, i.e. it can go up quickly during an storm and also can go down even before end of the storm, while the open channels have probably larger time-lags with rainfall.

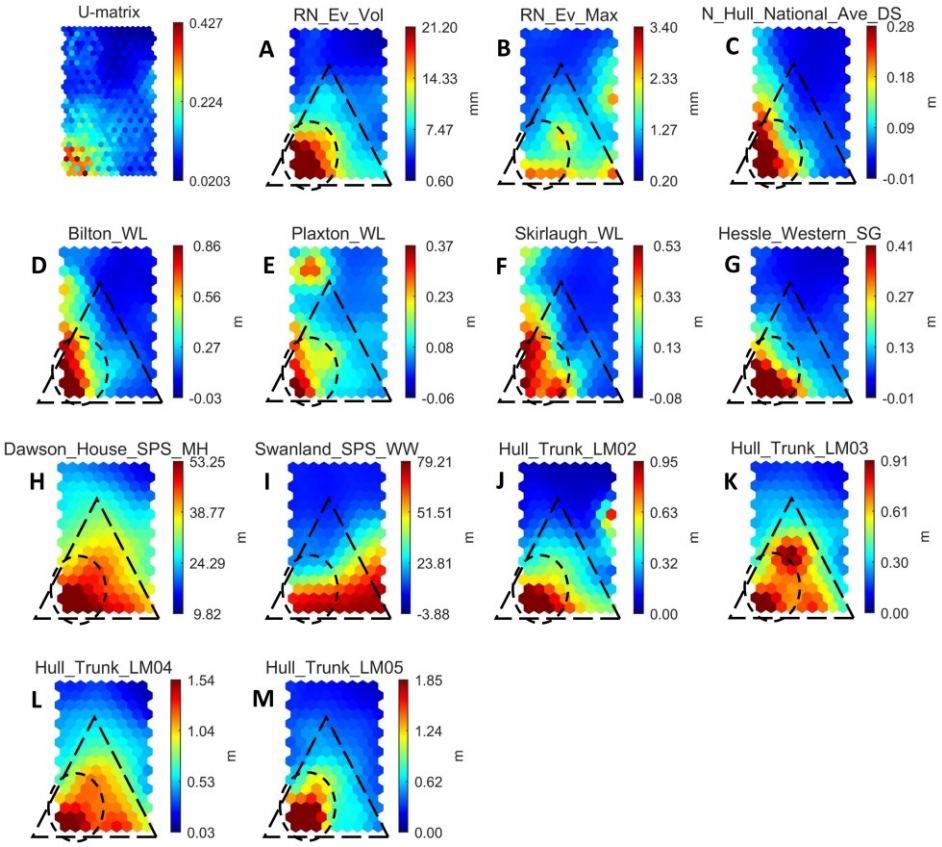

Figure 8: Relationships between rainfall in Cottingham (Lattices A and B: total precipitation and maximum intensity, respectively), water level in open channels (Lattices C to G: TS3, TS7, TS8, TS9 and TS2 in Table 1), and water level in sewer network (Lattices H to M: TS12, TS11, TS13, TS14, TS15 and TS16) for the data of a period of 10 years, from 1 Jan 2010 to 31 Dec 2019 (771 rainfall events).

To examine the possible time lags, therefore, another test was carried out. Since SOM is a method for visualising relationships in the data, not quantifying them, this test is done only as

a preliminary investigation of time lags between water levels and rainfall, an issue which is the objective of the predictive modelling in the Section 7.

In this test, in contrast to the last one, SOMs are not developed based on storm events. Rather than parameters specific to a storm such as maximum water level during a storm event, all data points in the water level time-series are used in the training of the SOMs. To include rainfall parameters with a lag, an 'observation window' is employed over which total precipitation is calculated. This window, for each point in the time-series, extends from the time of that point to a few hours before it. Figure 9 shows an example of water level profile (blue line) and three observation windows for calculating rainfall parameters in the last 2, 12 and 24 hours before a specific point in the water level time-series.

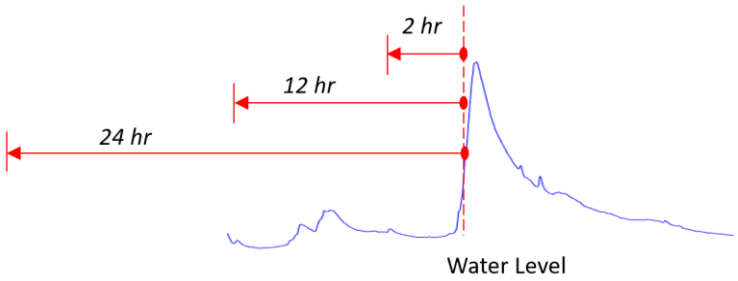

Figure 9: Design of SOMs based on data points with rainfall observation windows to take time lag between rainfall and water level into account.

Figure 10 shows the result of the test with water levels in four open channel watercourses (Lattices J to M) and four locations in the sewer network (Lattices F to I), with total precipitation in Cottingham at the same time as well as in the last 2, 12, 24 and 36 hours (Lattices A to E, respectively). Looking at Lattices A to E, the high rainfall cluster rotates around the bottom right corner with increasing the size of observation window. The dashed circles and red arrows in the figure indicate this issue. By comparing the shape and direction of high water level clusters in water level maps with the rainfall maps we can find out about the most possible time lags between the two. This comparison shows that the sewer network levels correlate with the total precipitation with observations window of ~2 hours and open channel levels with those of ~2-12 hours. This suggests that the time lag between water level in the sewer network and rainfall in Cottingham is probably less than 2 hours, while it could be between 2 and several hours in the open channel watercourses.

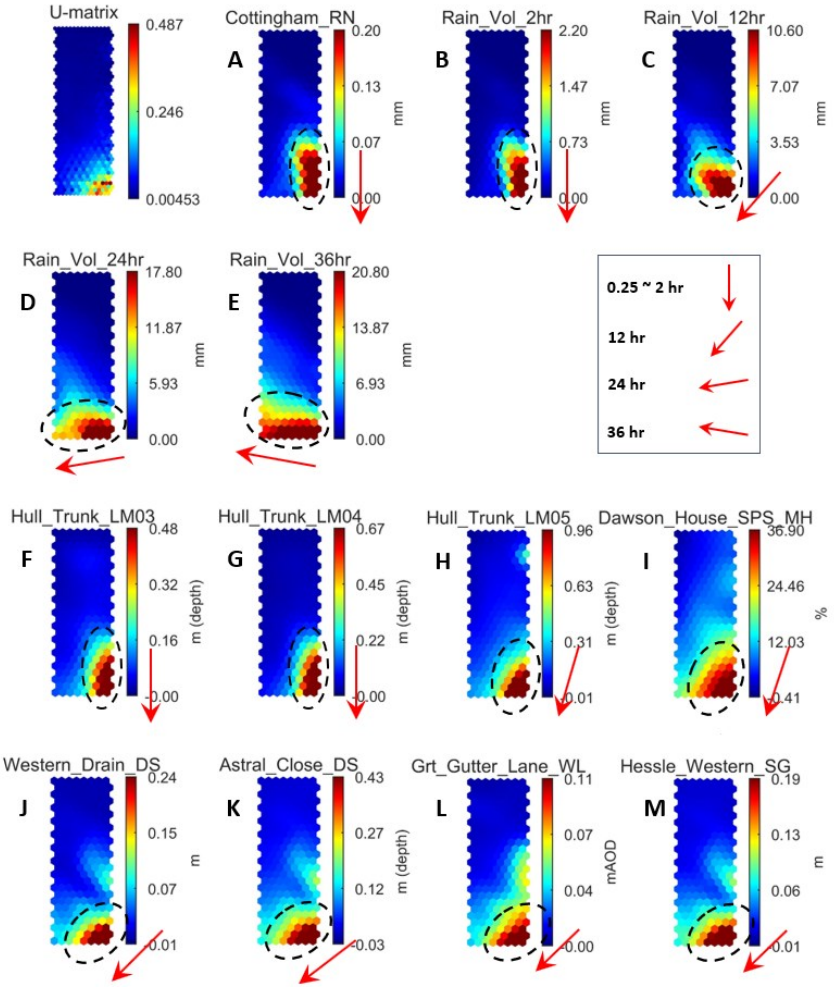

Figure 10: SOMs for exploration of time lags between total precipitation (Lattices A to E) and water level in the sewer network (Lattices F to I) and the open channels (Lattices J to M).

6.3 Summary of SOM analysis

In addition to the tests presented above, many other tests were performed using SOM which are not presented in this report. The major findings of all the tests are summerised as below.

- Rainfall in Cottingham shows a strong correlation with water level in both the open channel watercourses and the sewer network, but the linkage with the latter is more significant.
- Between rainfall parameters (total precipitation and maximum intensity), the former shows a stronger linkage with water level in the system.
- Deseasonalisation of water level time-series improves the correlations indicated by SOMs.
- The best dry period between two successive rainfall events in the automated rainfall event detection model is 6 hours for the present data.

- Response of sewer network to rainfall is much faster than open channels (below 2 hours for sewer network and between 2 and 12 hours for open channels).

7. Development of Predictive Models

The aim of this section is to quantify the relationships identified by the SOMs and proposing a methodology for water level forecasting based on historical data of rainfall, water level, and groundwater level. One of the identified relationships was about the difference between response of water level in the sewer network and open channels to rainfall. Therefore, the proposed methodology should be tested at least for one open channel watercourse and one trunk sewer. The SOMs analyses showed that water level in the trunk sewers responds much faster to rainfall. On the other hand, many of the open channel watercourses flow into the sewer network. For example, Setting Dyke flows to the sewer network at upstream of LM03, and there is no sewer monitor downstream of this connection. These two locations are selected as case studies for the purpose of testing the proposed predictive modelling approach. Figure 11 shows the location of these two measuring points as well as rainfall gauge at Cottingham, and the associated time-series. By using these two case studies, the following issues will be investigated/addressed.

- What time lag exists between rainfall at Cottingham and water level at Setting Dyke and LM03.
- How many hours in the future water level in the sewer network and open channels can be forecasted with a reasonable accuracy.
- Apart from rainfall, inclusion of what other parameters can enhance the predictions of water level in the sewer network (how and to what extent use of upstream open channel water level and groundwater level may lead to better forecasts).
- SOMs showed that among rainfall parameters, total precipitation is more significantly linked to water level in the system. But how far in the past total precipitation should be used to generate the most accurate forecast of water level in the future.

As discussed in Section 5, data-driven models are developed here based on three supervised ML techniques, namely ‘Regression FF’, ‘Classification RF’ and ‘Regression RF’ techniques, to address the above questions as well as propose a water level forecasting methodology.

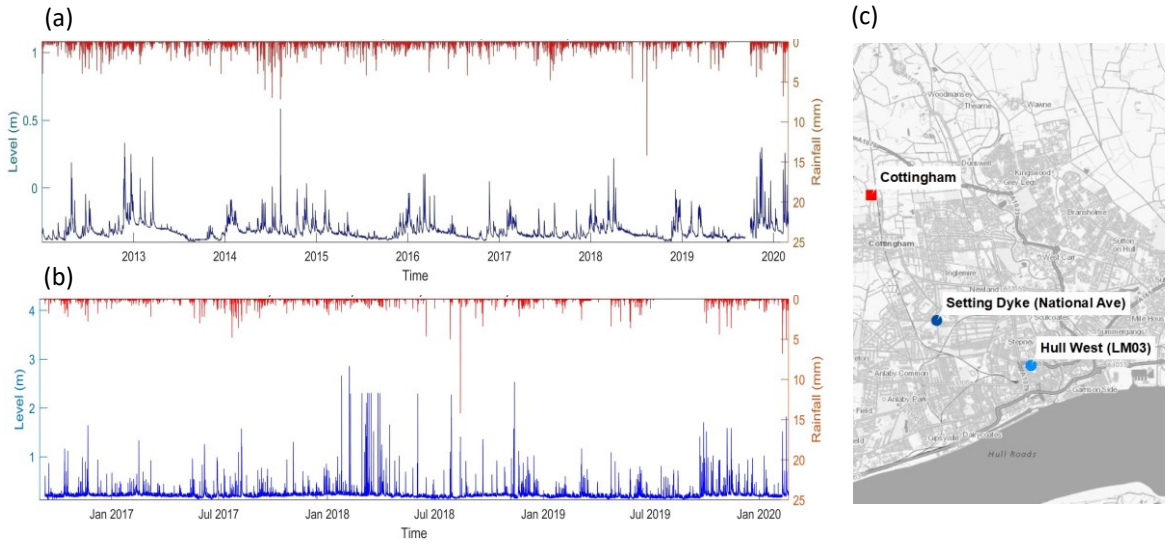

Figure 11: Water level time-series (blue lines) of the two test cases, a) Setting Dyke (TS3) and b) LM03 (TS14) with rainfall at Cottingham (red line); and c) all the three locations on the map.

7.1 Model design

Output parameter, to be forecasted, is water level in a location where we want to see whether flooding occurs, and input parameters could be a range of different parameters such as rainfall, water level in the same or other locations, groundwater level, etc, in the past. Figure 12 shows schematically the definition of input and output parameters for the predictive models. Note that in the water level predictions in this section, water level itself is taken into consideration rather than water level above baseline as in the event-based SOMs in Section 6.

Assume that we are at time t, and aim at predicting water level in a specific location (an open channel stream or the sewer network) T_F hours in the future, i.e. at $t + T_F$. In this case, output parameter is water level at $t + T_F$ and input parameters are rainfall, water level, gradient of water level profile, etc, in the last several hours. Therefore, 'observation windows' are defined for each input parameter over which the variable is determined. For example, total precipitation is calculated over the last T_R hours, i.e. between $t - T_R$ and t. For other input variables different sizes of observation windows are considered: T_M hours for mean water level and T_S hours for the gradient of the water level profile.

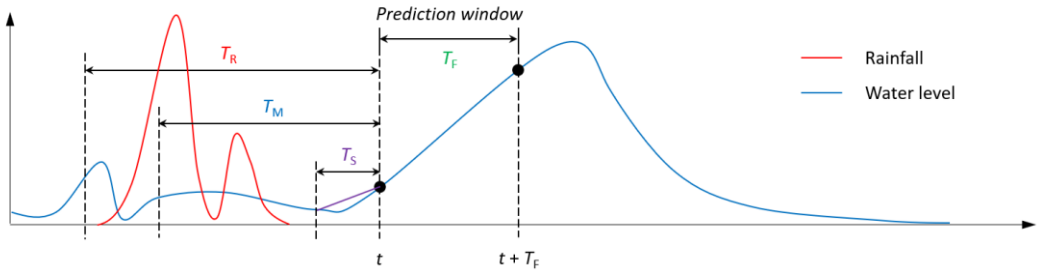

Figure 12: Input and output parameters for predictive modelling; prediction window (T_F) for the output parameter (water level) and observation windows for input parameters (T_R, T_M and T_S for rainfall, mean water level and gradient of water level profile, respectively).

The ML algorithms are used to train the model based on historical data, i.e. where both input and output parameters are available, and then the trained model is used to predict the output parameter for unseen (future) data. For example, suppose that rainfall and water level data are available for the past 10 days with a frequency of one hour, i.e. we have 240 data points available, and we want to use them to predict water level in the next 2 hours. Firstly, the model is trained based on the available 240 data points with for example input parameters such as total precipitation in the last 10 hours and mean water level in the last 5 hours; and then the model is used to predict the output parameter (water level) 2 hours ahead. In this example, in the training process, at each data point, total precipitation is calculated from that point to 10 points before it (T_R of 10 hours), mean water level is calculated between that point and 5 points before it (T_M of 5 hours), and the output parameter is set to the value of water level at two points ahead (T_F of 2 hours). The model is trained using this setup and after training it will be used for prediction of water level in the future. Therefore, it can be used as a warning system that gets values of parameters in the past several hours and gives predictions of water level in the next few hours.

The difference between regression- and classification-based approaches in the present application is that the output in the former is actual values of water level while in the latter it is classes of water level above or below a threshold. For instance, see Figure 13. It shows an example of prediction of 50 days of water level in Setting Dyke (TS3) in 2018 based on training of the model using data of about 8 years from 2012 to 2020 with classification and regression RF models. The data from 1 Jan 2012 to 30 June 2020 was split into two parts, 1) a period of 50 days from 1 Mar to 20 Apr 2020 put aside as unseen data to be predicted by model after training, and 2) the rest of the data for training the model. Regression RF model predicts actual values of water level as shown by red line in Figure 13-left (Regression FF model does the same). Classification RF model, instead, predicts whether water level goes above a threshold (= -0.1 m in this example) or not, as shown by coloured dots in Figure 13-right. These colour

dots show the predictions in classes of Positive and Negative, meaning whether the actual water level is above or below the threshold; and True and False, meaning whether they are predicted correctly by the model or not. Therefore, the output is divided into four categories of True Positive, True Negative, False Positive and False Negative predictions. For example, True Negative means actual value of observational data (water level) is below the threshold (Negative) and the model predicted it correctly (True).

In order to evaluate the performance of the models and estimate the accuracy of predictions, Root Mean Square Error (RMSE) and Nash-Sutcliffe Model Efficiency Coefficient (NSE) are employed as performance metrics for the Regression FF and RF models; and True Positive Rate (TRF), False Discovery Rate (FDR), and Matthews correlation coefficient (MCC) are applied for the classification RF model. For the definition of the metrics, see Appendix B.

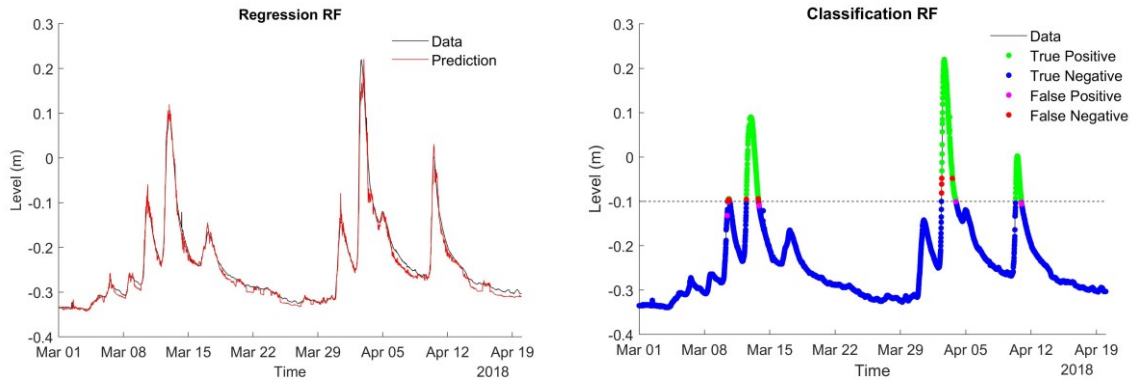

Figure 13: an example of the predictions of water level in Setting Dyke (TS3 in Table 1) by the Regression RF model (right) and Classification RF model (left), either to predict actual values of water level (right), or whether it goes above a threshold (-0.1 m) or not (left).

The forecast horizon, which is the length of time into the future at which model can provide forecasts, depends on several factors such as the time-lag exists in the actual data between different components of the system, for example between rainfall in Cottingham and water level in the trunk sewers in Hull; or other factors such as the combination of input parameters we use in the model, or the quality of the historical data used to train the model; or the performance of the ML techniques. One of the main objectives of this study was to investigate these issues for the data of LWWP organisations in Hull. Therefore, in order to examine these, test the model, and demonstrate how the model can be applied for water level forecasts, modelling the two test cases of Setting Dyke (TS3) and LM03 (TS14) are performed and the results are presented in the following sections.

7.2 Test case 1: Setting Dyke (TS3)

Eight years of data was employed to train the model and then it was used to predict water level in Setting Dyke for a period of about two weeks in March 2018 for a forecast horizon of 3 hours. For this analysis, three input parameters were considered: total precipitation in Cottingham in the last T_R hours, gradient of Setting Dyke's water level profile in the last T_S hours, and mean water level in Setting Dyke in the last T_M hours. Five combinations of these input parameters were used to train the model: A) total precipitation, B) total precipitation and gradient of water level profile, C) total precipitation, gradient of water level profile and mean water level, D) total precipitation and mean water level, and E) mean water level only. The input parameter combination with the highest performance metric values was considered to be the best option.

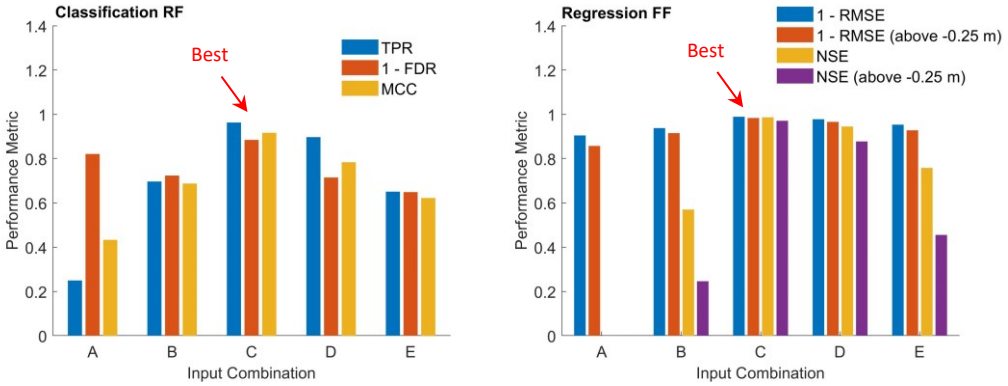

Figure 14: Performance metrics calculated for Classification RF (left) and Regression FF (right) predictions of Setting Dyke (TS3) water level when different combinations of input parameters are employed.

Figure 14 presents the calculated performance metrics for the combinations of input parameters A to E using Classification RF (left) and Regression FF (right) models. According to this comparison, the most accurate prediction is when combination C, i.e. total precipitation, gradient of Setting Dyke's water level profile and mean water level in the last several hours, are employed as input parameters. Figure 15 shows the predictions when combinations A, C and E are employed. When only rainfall is used (Figure 15-left), the time when water level starts rising in the prediction matches well with the data, but the value of water level is underestimated. When the input parameter is only mean water level (Figure 15-right), the predicted profile is smoother and the underestimation is small, but there is a lag of about 2~3 hours between prediction and data, that means the model is actually not fulfilling the forecast of 3 hours. However, when combination C is employed, there is a good agreement between prediction and data, which means the model can forecast water level 3 hours in the future with

a good accuracy using past values of rainfall in Cottingham and water level in Setting Dyke itself.

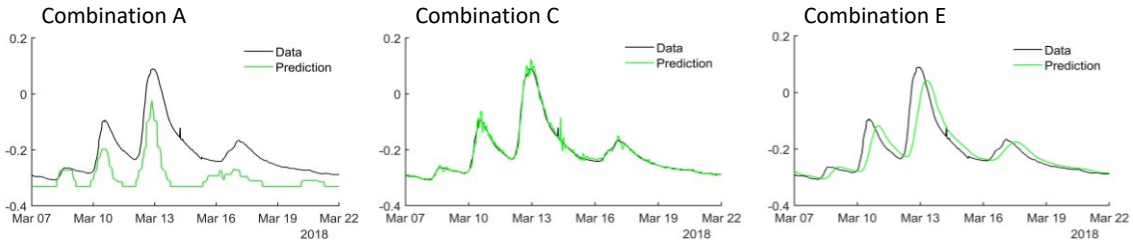

Figure 15: Predictions made by Regression FF model for water level in Setting Dyke using different combinations of input parameters: Combination A (left): only rainfall; Combination C (middle): rainfall, mean water level, gradient of water level profile; and Combination E (right): mean water level only.

Sensitivity analyses of observation window size for rainfall, mean water level, and gradient of water level profile were carried out. The aim was to find the values of T_R, T_M and T_S which give the highest accuracy of predictions of Setting Dyke's water level for a forecast horizon of 3 hours. Figure 16 shows the result of analysis using Regression RF (left) and Regression FF (right) models for the rainfall observation window size. It reveals that the best value of T_R is 18 hours for both Regression RF and FF models. In other words, when total precipitation in the last 18 hours is used as input, the most accurate prediction for 3 hours in the future is achieved. Similar analyses were carried out for T_M and T_S, and results showed that the best values for these two observation windows for Setting Dyke are 6~9 hours and 2 hours, respectively (the results are not presented here).

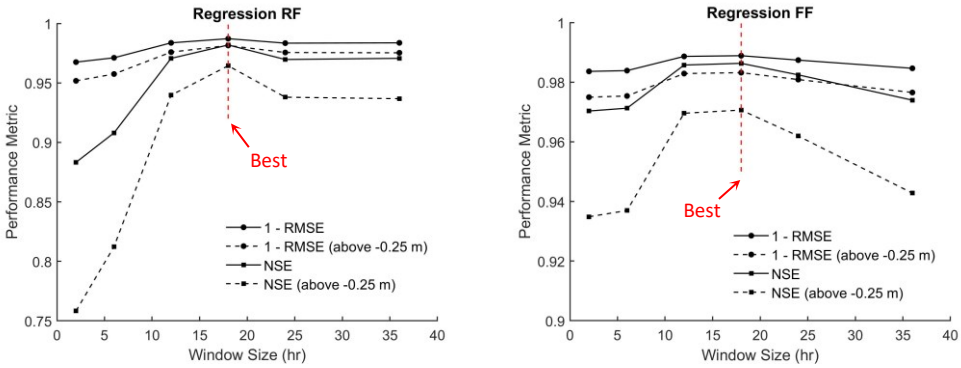

Figure 16: Accuracy of predictions of water level in Setting Dyke three hours in the future using different sizes of rainfall observation window, T_R, with Regression RF (left) and Regression FF (right) models.

Sensitivity analysis was also done on the forecast horizon T_F and the results are presented in Figure 17. T_F values of 1 to 9 hours were tested, and the performance metrics show that, by increasing T_F, accuracy of predictions goes down, as expected. However, up to 3 hours, the

accuracy of predictions is high. This can also be seen from the profiles of predictions presented in Figure 18. It can be concluded that the model can predict water level in Setting Dyke up to 3 hours in the future with a good accuracy; and allowing for some uncertainties, it can predict it up to 4 hours.

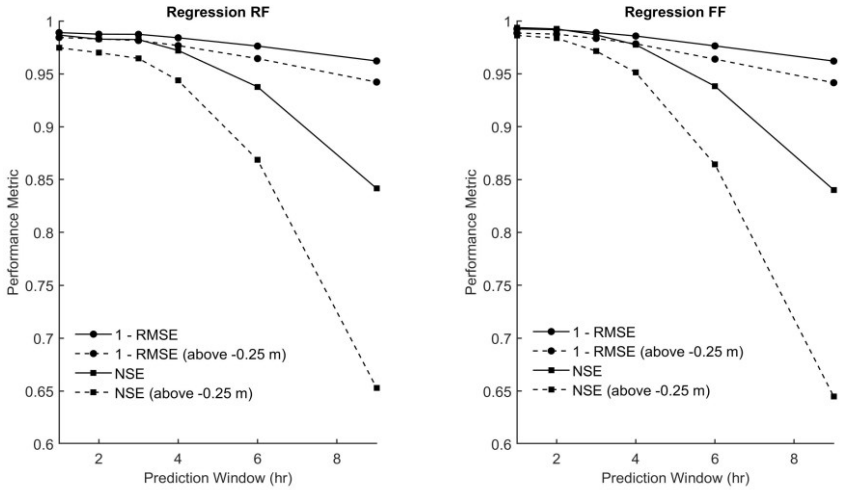

Figure 17: Accuracy of predictions of water level at Setting Dyke for different forecast horizons T_F using Regression RF (left) and Regression FF (right) models.

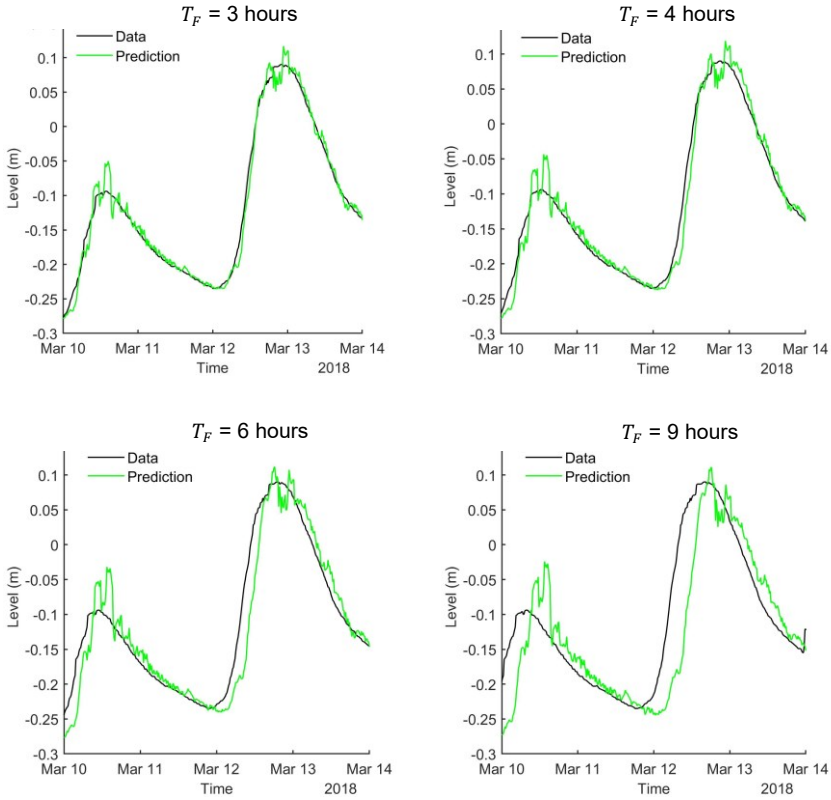

Figure 18: Setting Dyke water level prediction using Regression FF model for different forecast horizons, T_F. Black: actual data; green: prediction.

7.3 Test case 2: LM03 trunk sewer (TS14)

From the SOM analysis in Section 6, we expect a smaller forecast horizon for trunk sewers in Hull than open channel watercourses. For this test, forecast horizon T_F is set to 45 minutes and then sensitivity analyses on the size of observation windows of total precipitation at Cottingham (T_R) and gradient of LM03's water level profile (T_S) are performed. The most accurate predictions were achieved using T_R and T_S values of 6~8 and 1~2 hours (depending on the LM model applied for predictions), respectively.

The inclusion of additional input parameters, such as mean water level at LM03 itself or at upstream open channels, and groundwater level is discussed in the following.

Figure 19 presents predictions of water level at LM03 45 minutes in the future using two different combinations of input parameters using the Regression FF model. The left figure is when past values of total precipitation at Cottingham and the gradient of LM03's water level profile are employed as input parameters to train the model, and the right figure shows the prediction when mean water level is also included. Adding mean water level makes the output smoother and closer to the observed data at higher values, but it creates a time lag between prediction and data, i.e. slightly reduces the prediction window (by about 15 minutes in this test). This is probably because the change of water level in the sewer network is quite fast. Therefore, mean water level at LM03 is not considered as an input for water level predictions in LM03.

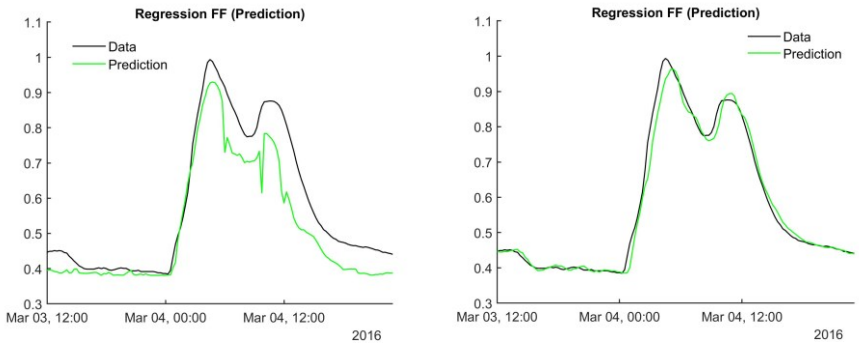

Figure 19: prediction of LM03's water level using Regression FF model, 45 minutes in the future; left: with input parameters of total precipitation at Cottingham and gradient of LM03's water level profile; and right: including mean water level at LM03 as well.

In a new test, mean water level at Setting Dyke in the last 6 hours, as an open channel upstream to LM03, is employed as an additional input parameter for the prediction of water level at LM03. Figure 20 presents the results of predictions using Regression RF model with and without Setting Dyke's mean water level as input. As seen, especially in the graphs (c) and (d) where both cases are compared, when mean water level at Setting Dyke is added

(blue line), not only the falling limb of the event is predicted with a higher accuracy, but also the rising limb, where water level starts rising, is also estimated more accurately, resulting in a smaller gap between prediction and data. This has improved the forecast horizon by about 15 minutes.

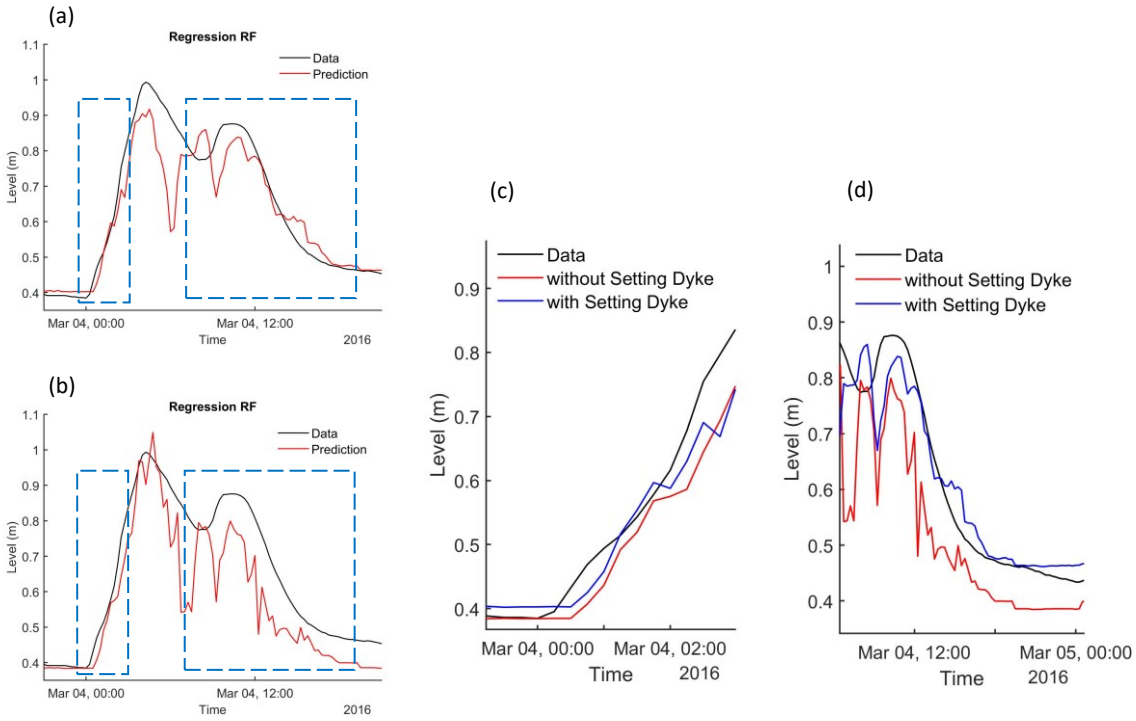

Figure 20: prediction of LM03's water level 45 minutes in the future using Regression RF model with and without mean water level at Setting Dyke as an input parameter. a) input parameters: total precipitation at Cottingham and gradient of LM03's water level profile; b) the same parameters plus Setting Dyke's mean water level; c) and d) comparison of profiles in the rising and falling limbs of the event shown by dashed lines in plots (a) and (b).

Two tests were performed to examine the inclusion of groundwater level data as input for prediction of water level in LM03. Mean groundwater level data at Cottingham Willerby Hill and Cottingham North Houses (TS4 and TS5, respectively, in Table 1) were used and compared. In this case, input paramters are total precipitation at Cottingham, gradient of LM03's water level profile, and groundwater level at Willerby Hill in one test, and groundwater level at North Houses in another test, to predict water level 45 minutes in the future. The results are presented in Figure 21. This comparison shows that no improvement is achieved as a result of adding North Houses groundwater level, but Willerby Hill groundwater level improves the forecast horizon by around 10-15 mins.

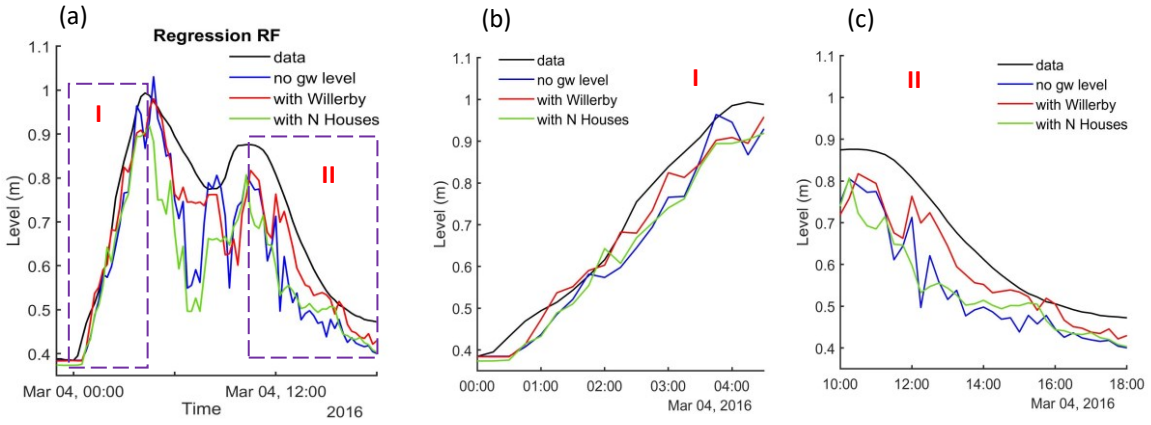

Figure 21: a) comparison of predictions of LM03 using Regression RF model with and without groundwater level data; b) and c) the rising and falling limbs of the event marked by I and II. Black line: data; blue line: no groundwater level data; red line: with Willerby Hill groundwater level (TS4); and green line: with North Houses groundwater level (TS5).

A sensitivity analysis was done on the size of prediction window (forecast horizon) to see how far in the future the water level at LM03 can be predicted. Based on the analyses presented above, input parameters were set to total precipitation at Cottingham in the last 6 hours, gradient of the water level profile in LM03 in the last 1.5 hours, mean water level at Setting Dyke in the last 6 hours, and mean groundwater level at Willerby Hill in the last 1 hour, to predict water level in LM03 with different sizes of prediction window T_F.

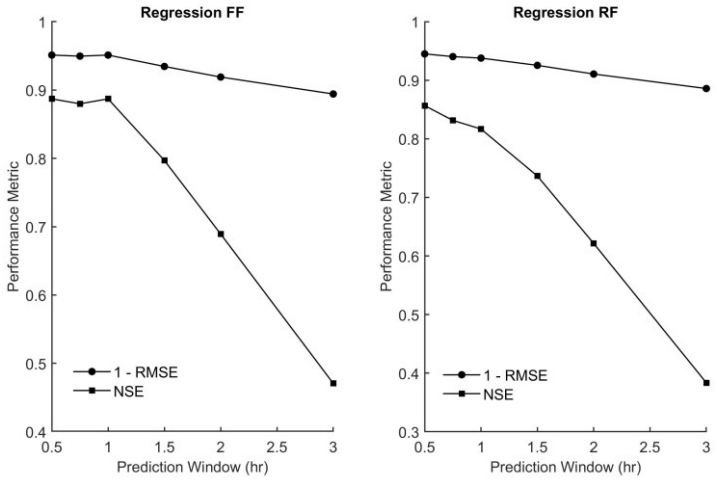

Figure 22: Accuracy of predictions of LM03's water level using different sizes of prediction window T_F (forecast horizon) using Regression FF (left) and Regression RF (right) models.

The result is presented in Figures 22 and 23, where performance metrics for Regression RF and FF models, and predictions using different T_F values by the RF model are shown, respectively. As expected, by increasing T_F, accuracy of prediction decreases. However, this decrease occurs very slowly when the forecast horizon is below one hour. It is concluded that

the model can forecast water level in LM03 up to 45 mins in the future with a good accuracy, and allowing for some uncertainty, it can predict it up to 1 hr.

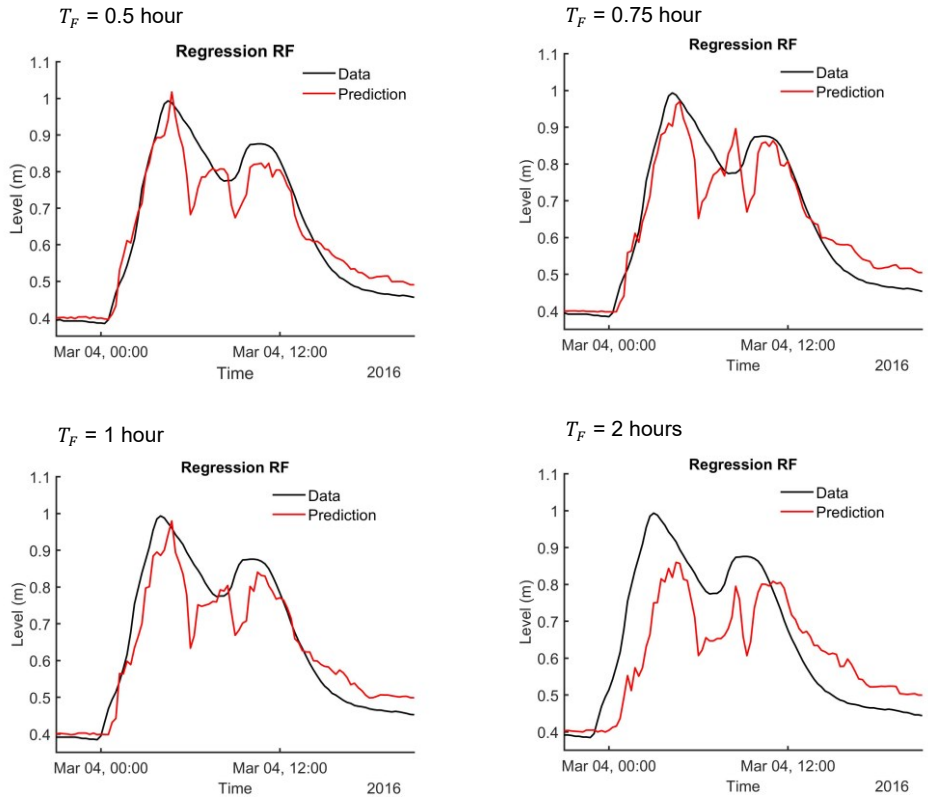

Figure 23: Predictions of LM03's water level profile using Regression RF model with different sizes of prediction window T_F.

7.4 Summary of the predictive models

Predictive models were developed based on both classification-based and regression-based ML algorithms. The model was tested for two case studies: prediction of water level in an open channel watercourse (Setting Dyke) and a trunk sewer (LM03) in Hull. Various combinations of input parameters such as rainfall, gradient of water level profile, mean water level in the same and/or another location, and groundwater level data, were tested. Sensitivity analyses on the sizes of observation windows for these parameters, i.e. the period in the past over which they should be determined as input for the predictions, were performed; and finally, the best prediction windows, i.e. how far in the future water level in these two watercourses can be predicted were estimated. The following points are concluded:

- The model can generate forecasts for individual locations based upon historical rainfall, water level, slope of water level change, and groundwater level data.
- The farthest in the future water level can be predicted with a good accuracy is 3~4 hours for Setting Dyke (open channel) and 45~60 minutes for LM03 (sewer).

- The best combination of investigated input parameters for water level forecasts in Setting Dyke are rainfall at Cottingham, and mean water level and gradient of water level profile in Setting Dyke itself. The best observation windows for these parameters, i.e. T_R, T_M and T_S are 18, 6~9 and 2 hours, respectively.
- The best combination of investigated input parameters for water level forecasts in LM03 are rainfall at Cottingham, gradient of water level profile in LM03 itself, mean water level profile in Setting Dyke, and groundwater level in Willerby Hill. The best observation windows for these parameters, are 6, 1.5, 6, and 1 hour, respectively.

8. Implementation of Model

The developed predictive models can be applied as an early warning tool for the two locations tested in this study, i.e. Setting Dyke and LM03. The models should be trained based on historical data, and then fed into by values of rainfall, water levels, and groundwater levels in the preceding several hours to predict water level 3 hours and 45 minutes in the future, respectively.

The model codes (in MATLAB R2019b) are provided along with relevant User Guides. The codes include the model for filtering data (as described in Section 4), the SOM model for exploration of relationships (as described in Section 6), and the predictive models (described in Section 7). The input/output set-up for the forecasts in Setting Dyke and LM03 should be defined based on the sensitivity analyses performed in Section 7.2 and 7.3. For real-time applications, the model should be fed into by relevant data in real-time. If there is a change in the system, the model should be trained and calibrated for the new condition.

To apply the model for other locations, such as other trunk sewers or open channels, the model should be trained based on their own historical data. Then, it can be used for prediction of water level in the new locations.

For replication of the approach for other regions, the model should be set up and trained based upon their local system data; sensitivity analyses on input/output parameters, ML hyperparameters, size of observation windows and prediction windows should be carried out; and the model should be tested with relevant case studies before being used for water level forecasts.

To improve efficiency and applicability of the model, a flood risk model can be incorporated into the model to be able to estimate risk of exceedance of water level triggers. This risk calculation could be done by predicting exceedance probabilities for a range of rainfall values

and combining those probabilities with the rate of occurrence of the critical rainfall (e.g., rainfall corresponding to flooding with a probability of 60%).

9. Overall Summary and Conclusions

A large amount of data of rainfall, water level and groundwater level across Hull and the surrounding area was collected from the project partners (HCC, YW, EA, and ERYC) to study the relationships between different elements of the drainage system in Hull. The data was cleansed and combined into a single dataset. ML algorithms were used to explore the relationships within the data and develop predictive models which can be employed as a warning tool. Predictive models forecast water level in the future using past values of rainfall, water level and groundwater level. The model was tested for two locations, Setting Dyke (open channel) and LM03 (trunk sewer). The results of analysis showed that the farthest in the future water level can be predicted with a good accuracy is 3~4 hours for Setting Dyke and 45~60 minutes for LM03. The project, by the data analyses performed as well as the collaborations with the project partners through the data collection process and meetings and workshops, had the following outcomes.

- A better understanding of the existing telemetry network across the study area.
- Clarity on the variability of quality of data and physical parameters.
- Identification of important relationships between network elements.
- Better use of current systems and evidence to support future funding investments.
- Improved flood resilience to the area through development of an early warning tool.
- Demonstration of the value of combining and sharing data among the different LWWP partners, along with the value of data-driven methods to help understand the behaviour of complex systems.

The output of the project is a ML water level predictive system which can be used as an early warning tool for predicting water level exceedance above defined thresholds. The model can be used for water level forecasts in Setting Dyke and LM03. The approach can be replicated for other areas given that the historical data is available, and the model is trained and tested for that system.

Telemetry data in Hull (and many other similar areas in the country) are collected and stored by water organisations who use it for their individual use, with limited sharing of data between them. By collecting data in a more systematic way, and incorporating ‘flood risk’ modelling into the system in the future studies, it could then be used for more informed decision makings.

For short-term operational decisions, trigger values of total precipitation and water levels at specific locations could be used to identify immediate risks to the system. For strategic level decisions, a large area could be assessed for overall risk of water level exceedance above defined thresholds under typical high-rainfall conditions, supporting maintenance prioritisation and related decisions.

References

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H., (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412-424.

Kind, M.C., Brunner, R.J. (2013). SOMz: photometric redshift PDFs with self organizing maps and random atlas. Monthly Notices of the Royal Astronomical Society, 438(4), 3409-3421.

Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J. (1996). Engineering applications of the self-organizing map. Proceedings of the IEEE, 84, 1358-1384.

Meyers, G., Kapelan, Z., Keedwell, E. (2017). Short-term forecasting of turbidity in trunk main networks. Water Res., 124, 67-76.

Speight, V., Mounce, S.R., Boxall, J.B. (2019). Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. Environ. Sci.: Water Res. Technol., 5, 747-755.

Appendix A

A list of data provided by the project partners from monitoring stations across Hull and East Riding of Yorkshire is presented in Table A below. Figure A shows the location of the stations. Note that not all the stations are shown on the map since coordinates of some of them are not available.

Table A: a full list of data provided by project partners.

No	Organisation	Location/Tag	Type	Easting	Northing
1	EA	Anlaby	Groundwater level	503305	427903
2	EA	Bentley	Groundwater level	501895	435936
3	EA	Cherry Tree Cottage	Groundwater level	504346	438565
4	EA	Cottingham North Houses	Groundwater level	505000	435000
5	EA	Cottingham Willerby Hill	Groundwater level	502281	431806
6	EA	Dunswell	Groundwater level	507500	435500
7	EA	East View	Groundwater level	505300	434400
8	EA	Ella Cross Roads	Groundwater level	500500	430200
9	EA	Ennerdale	Groundwater level	508191	434303
10	EA	Hampston Hill	Groundwater level	505400	437000
11	EA	Harland Rise	Groundwater level	503100	433600
12	EA	Hollycroft	Groundwater level	504300	437400
13	EA	Ideal Standard	Groundwater level	506747	429952
14	EA	Kenley Reach Thearne	Groundwater level	507285	437161
15	EA	Little Weighton	Groundwater level	498728	433854
16	EA	Meeting House Cott	Groundwater level	505200	433500
17	EA	Mount Pleasant	Groundwater level	500595	440113
18	EA	N Houses Cottingham	Groundwater level	505165	435151
19	EA	N Moor House Chalk	Groundwater level	504900	435000
20	EA	North Houses Drift	Groundwater level	505165	435151
21	EA	Northlands	Groundwater level	498661	438281
22	EA	Poplar Farm	Groundwater level	503600	435600
23	EA	Ralph Nook	Groundwater level	499801	433899
24	EA	Sunnydene Plaxton Br	Groundwater level	506461	436432
25	EA	Swift Caravans	Groundwater level	505100	433900
26	EA	Walkington	Groundwater level	499500	437800
27	EA	Walkington Wold	Groundwater level	497540	436351
28	EA	Westwood	Groundwater level	502009	437969
29	EA	Willerby Haggs	Groundwater level	503240	431197
30	EA	Willerby Hill	Groundwater level	502281	431806
31	EA	Wood Farm Cottingham	Groundwater level	505500	434300
32	EA	Beverley Shipyard	Open channel water level	505397	439744
33	EA	Brough West Clough	Open channel water level	496296	424617
34	EA	Dunswell Ennerdale Bridge	Open channel water level	508191	434303

35	EA	East Hull Hedon Road	Open channel water level	513097	429300
36	EA	Hempholme Weir	Open channel water level	507957	449913
37	EA	Hessle Western Drain	Open channel water level	503295	427583
38	EA	Hull Barrier	Open channel water level	510194	428354
39	EA	Hull High Flags	Open channel water level	509998	429679
40	EA	Setting Dyke (Birdsall Avenue)	Open channel water level	505013	430528
41	EA	Setting Dyke (National Avenue)	Open channel water level	506692	430535
42	EA	Paull	Open channel water level	516516	426331
43	EA	Scurf Dyke	Open channel water level	507851	450592
44	EA	Stone Ferry Bridge	Open channel water level	510153	431427
45	EA	Wilfholme Ps Barmston	Open channel water level	506211	447044
46	EA	Cottingham	Rainfall	504791	434188
47	ERYC	Acre Head Drain	Open channel water level	504260	427591
48	ERYC	Acre Heads Drain Hull Rd	Open channel water level	503978	428781
49	ERYC	Astral Close Screen	Open channel water level	503670	427320
50	ERYC	Atwick Village Drain	Open channel water level	518940	450847
51	ERYC	Bh5 Washdyke Bridge	Open channel water level	514513	446063
52	ERYC	Bilton	Open channel water level	517221	432531
53	ERYC	Bond Street	Open channel water level	520005	428424
54	ERYC	Bowlams Dike	Open channel water level	510797	444740
55	ERYC	Broomfleet	Open channel water level	487449	425601
56	ERYC	Burstwick Drain - Hedon	Open channel water level	518722	428074
57	ERYC	Carr Lane	Open channel water level	504525	430544
58	ERYC	Cascade	Open channel water level	506764	430530
59	ERYC	Dutch River	Open channel water level		
60	ERYC	Filling Station	Open channel water level	501883	431012
61	ERYC	Fleet Drain	Open channel water level	503788	426223
62	ERYC	Great Gutter Lane	Open channel water level	502128	430615
63	ERYC	Hessle Haven	Open channel water level	503331	425829
64	ERYC	Hilderthorpe Screen	Open channel water level	517840	466762
65	ERYC	Hollym Screen	Open channel water level	534642	425048
66	ERYC	Hook Drain Goole	Open channel water level		
67	ERYC	Hornsea Allotment Track	Open channel water level	519978	446713
68	ERYC	Hornsea Burton Road	Open channel water level		
69	ERYC	Hornsea Mere	Open channel water level	519838	447210
70	ERYC	Inmans Estate	Open channel water level		
71	ERYC	Meaux Bridge Hold Drain	Open channel water level		
72	ERYC	Monk Dike	Open channel water level		
73	ERYC	N Frodingham	Open channel water level		
74	ERYC	Nelson St Pier	Open channel water level	510003	428090
75	ERYC	Plaxton Bridge	Open channel water level	506611	436548
76	ERYC	Pocklington	Open channel water level	481162	450628
77	ERYC	Preston	Open channel water level		
78	ERYC	R Hull - Beverly Beck	Open channel water level	505725	439367
79	ERYC	Rawdale Lagoon	Open channel water level	500905	430522

80	ERYC	Reedness Village	Open channel water level		
81	ERYC	River Derwent	Open channel water level		
82	ERYC	River Foulness	Open channel water level		
83	ERYC	River Hull Beverly Beck	Open channel water level	505711	439360
84	ERYC	River Ouse	Open channel water level		
85	ERYC	Robson Cottage Lagoon	Open channel water level	501682	430683
86	ERYC	Rudston	Open channel water level		
87	ERYC	Skirlaugh	Open channel water level	514217	439770
88	ERYC	South Cave	Open channel water level		
89	ERYC	Stone Creek	Open channel water level		
90	ERYC	Thornham Close	Open channel water level	491669	430868
91	ERYC	Tranby Lagoon	Open channel water level	504220	427780
92	ERYC	Wassand Estate	Open channel water level	517190	446531
93	ERYC	Well Lane	Open channel water level	503450	430685
94	ERYC	West of Seaton	Open channel water level		
95	ERYC	Western Drain Culvert	Open channel water level	503511	427633
96	ERYC	Willy Howe Wold Newton	Open channel water level	507137	472524
97	ERYC	Albion Mills - Willerby	Rainfall		
98	ERYC	Bridlington	Rainfall	518322	469291
99	ERYC	Brough	Rainfall	493552	428361
100	ERYC	Brough Rain Gauge	Rainfall	493552	428361
101	ERYC	Driffield Showground	Rainfall	501928	456714
102	ERYC	Elloughton	Rainfall		
103	ERYC	Goole	Rainfall		
104	ERYC	Hedon	Rainfall		
105	ERYC	Hessle Rain Gauge	Rainfall		
106	ERYC	Lock Hill Goole	Rainfall	474703	423449
107	ERYC	Market Weighton	Rainfall	486735	441961
108	ERYC	Nafferton Rain Gauge	Rainfall	505537	459366
109	ERYC	Pocklington	Rainfall	481054	450650
110	ERYC	Tranby Lagoon	Rainfall	504220	427780
111	ERYC	Willow Grove Screen	Rainfall	502779	439805
112	ERYC	Willy Howe Wold Newton	Rainfall	507026	472544
113	ERYC	Withernsea Rain	Rainfall		
114	HCC	Cottingham	Open channel water level	506014	433953
115	HCC	Counter Dyke	Open channel water level	507374	434351
116	HCC	Hessle Road	Open channel water level	504504	426595
117	HCC	Sand Dyke	Open channel water level	504221	429705
118	HCC	North Bridge	Rainfall	510247	429253
119	YW	Beverley	Rainfall	505412	441904
120	YW	Beverley Rural (N)	Rainfall	492109	438985
121	YW	Beverley Rural (S)	Rainfall	498903	437192
122	YW	Bransholme	Rainfall	507447	434094
123	YW	Bridlington Rural	Rainfall	517030	463539
124	YW	Central (Haltemprice)	Rainfall	500229	430210

125	YW	Cottingham	Rainfall	506366	434055
126	YW	Driffield	Rainfall	501816	459364
127	YW	Driffield Rural	Rainfall	485096	457475
128	YW	Gilberdyke Rural	Rainfall	477570	436739
129	YW	Hedon	Rainfall	515579	427710
130	YW	Hornsea	Rainfall	520129	450189
131	YW	Hornsea Rural	Rainfall	507304	447572
132	YW	Hull East	Rainfall	508957	432835
133	YW	Hull West	Rainfall	508056	427051
134	YW	Market Weighton	Rainfall	484099	440606
135	YW	North Ferriby	Rainfall	496973	432725
136	YW	Pocklington Rural	Rainfall	483690	456832
137	YW	South Cave	Rainfall	492109	438985
138	YW	Withernsea	Rainfall	533173	429519
139	YW	Withernsea Rural	Rainfall	521236	434070
140	YW	Beverley Road	Sewer water level	508363	431688
141	YW	Boothferry Road Wet Well	Sewer water level	503452	427346
142	YW	Burma Drive Wet Well	Sewer water level	513527	429783
143	YW	Capstan Road	Sewer water level	508186	433761
144	YW	Cliff Bridge Wet Well	Sewer water level	502530	425550
145	YW	Compass Road Wet Well	Sewer water level	508556	433659
146	YW	Cottingham George St	Sewer water level	504349	433008
147	YW	Dawson House Manhole	Sewer water level	508363	431688
148	YW	Dunscombe Park Wet Well	Sewer water level	510828	430552
149	YW	Ferry Rd CSO	Sewer water level	503397	426056
150	YW	Hull Anlaby CSO	Sewer water level	503987	428763
151	YW	Lm01 Hull Central	Sewer water level	502909	428638
152	YW	Lm02 Hull West	Sewer water level	506563	431909
153	YW	Lm03 Hull West	Sewer water level	509415	429239
154	YW	Lm04 Hull East	Sewer water level	510791	429472
155	YW	Lm05 Hull East	Sewer water level	512969	430919
156	YW	Lm06 Hull East	Sewer water level	514082	431346
157	YW	Needlers Way Wet Well	Sewer water level	509416	430835
158	YW	Swanland	Sewer water level	500433	427678
159	YW	Tudor Court	Sewer water level	502333	430168

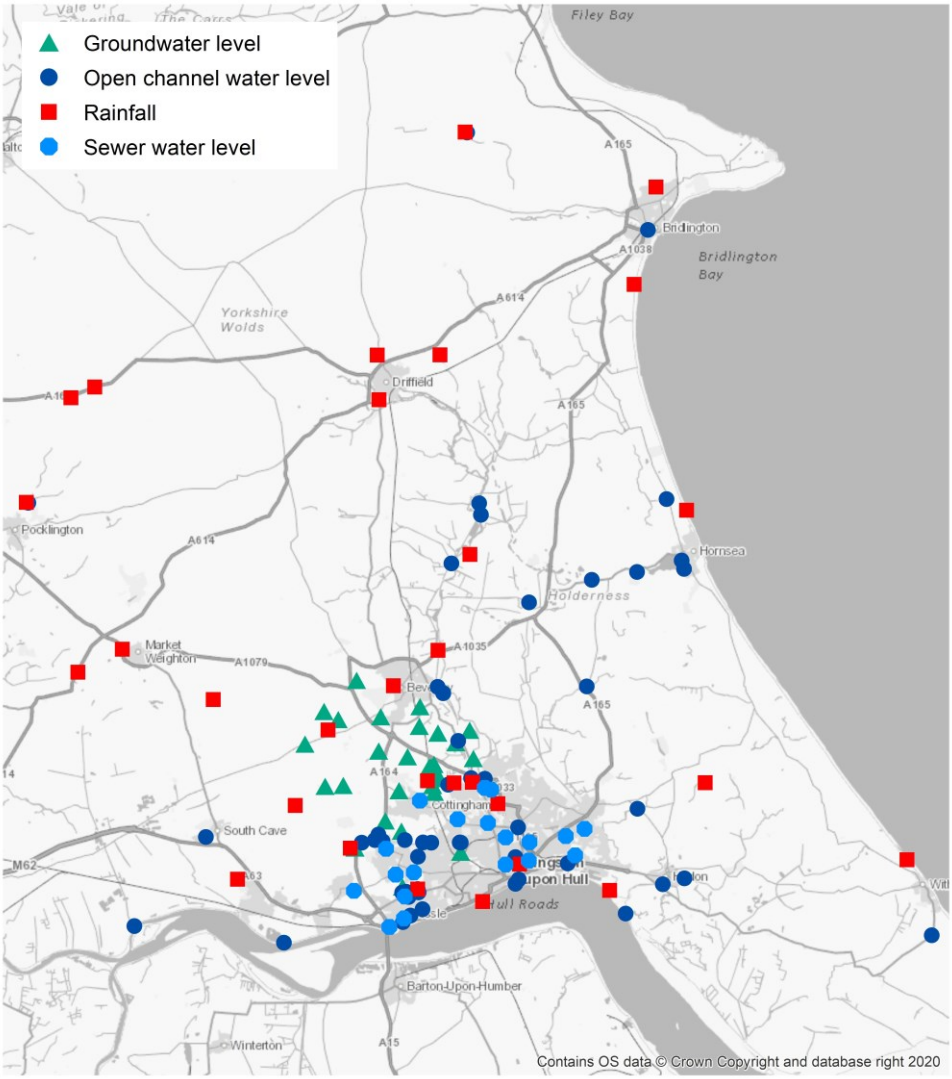

Figure A: location of monitoring stations across Hull and East Riding of Yorkshire for which data is listed in Table A.

Appendix B

The metrics used for assessing performance of the predictive algorithms in Section 7 are defined as in Equations (B1) to (B5). Root Mean Square Error (RMSE) and Nash-Sutcliffe Model Efficiency Coefficient (NSE) are employed for the Regression FF and RF models; and True Positive Rate (TRF), False Discovery Rate (FDR), and Matthews correlation coefficient (MCC) are applied for the classification RF model.

$$RMSE = \sqrt{\sum \frac{\left(Y_{pred} - Y_{data}\right)^2}{N}} \tag{B1}$$

$$NSE = 1 - \frac{\sum (Y_{Pred} - Y_{data})^2}{(\bar{Y}_{data} - Y_{data})^2} \tag{B2}$$

$$TPR = \frac{TP}{TP + FN} \tag{B3}$$

$$FDR = \frac{FP}{FP + TP} \tag{B4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{B5}$$

where Y_{data} and Y_{pred} are the actual (observed) and predicted values, respectively; $\bar{Y}_{data}$ is the mean of observational data; N is the number of data points in the part of time-series which is predicted; TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative, respectively. TPR represents the probability that the model will correctly predict positive class values (water level above threshold); FDR represents the probability a model predicts a positive when in reality no such event occurred, also known as a false alarm (Meyers et al., 2017); and MCC is a measure of summarising performance even when there is a skew in class sizes (Baldi et al., 2000). It is noted that in the analysis in Section 7.2 and 7.3, for the Classification RF model, $1 - FDR$, and for the Regression RF and FF models, $1 - RMSE$ are employed instead of FDR and $RMSE$ in order to be consistent with other metrics (i.e. having 1 for the best performance and 0 for the worst performance for all metrics), because perfect match with data corresponds to a FDR and $RMSE$ of 0, while to the value of 1 for other metrics. Also note that for the Regression RF and FF models, $1 - RMSE$ and NSE are calculated not only for the entire water level profile in the predicted section, but also for a subset of data above a threshold (e.g., -0.25 m in the tests in Section 7.1), because most part of the profile

is at a constant water level which is around the normal level of the channel. Only sometimes water level goes up due to rainfall. Therefore, using the part above a threshold above the normal level will give us a better metric of accuracy of the predictions.

Email: iCASP@leeds.ac.uk

Website: https://icasp.org.uk/

Twitter: @YorkshireiCASP