River flow accumulation and transit times in Bradford's river

Utkarsh Balooni
201792310

Supervised by: Onno Bokhove (School of Mathematics)
External advisors: Robert Hellawell, Water Quality officer (Aire Rivers Trust);
Prof Barney Lerner, Chair (Friends of Bradford's Becks)

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analitics
as part of the degree of

Master of Science in Data Science and Analystics

The University of Leeds, School of Mathematics

2 September 2024

UNIVERSITY OF LEEDS

School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

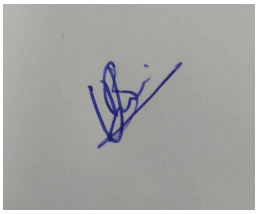
I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name Utkarsh Balooni

Student ID 201792310



Abstract

Urban river pollution is a pressing issue that threatens public health, degrades ecosystems and destroys natural beauty. This study aims to support the Friends of Bradford's Becks to better understand the catchment hydraulics of Bradford Beck, a small river system in Bradford, UK, plagued by pollution. The main objective is to assist in locating the source of the pollution by calculating flow accumulation along the river and transit times of pollution. Yorkshire Water provided flow and quality data spanning one year for strategic sites along the river. Exploratory data analysis revealed that the western part of the catchment is less urbanised, and the river discharge is highly variable. The percentage contribution to the total flow for each site is calculated using cumulative flow values for both dry and wet weather conditions. Flow accumulation at each site is then compared to the catchment areas to draw insights about runoff rates. The analysis revealed that the flow is highly 'flashy' for most of the catchment and changes rapidly. Also, catchment areas with more impervious, man-made structures seem to have higher surface runoff rates. The correlation of flow with land cover is also investigated and the effect of a diversion tunnel on flow accumulation is discussed. This study computes transit times by measuring the time difference between peak TAN (total ammonia nitrogen) values and filtering out significant water quality events that exceeded the given TAN threshold. The analysis identified May, June, and August as the months with the most pollution events. Pollution transit times and velocities for different river reaches were reported and each site's peak TAN value across all events was identified. Additionally, a complete Python framework for identifying TAN peaks and calculating transit times is provided to Friends of Bradford's Becks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Rivers have historically been the cornerstone of civilisation and have proven to be a valuable asset to humankind. They are also integral to maintaining ecological balance and sustaining biodiversity. Today, urban rivers face significant water quality challenges that reduce their recreational value, harm aquatic life, and deteriorate their aesthetic appeal. According to the UNESCO World Water Assessment Programme [2017], more than 80% of the world's wastewater is discharged into the environment without treatment. Moreover, the rise of industries and manufacturing plants releases significant amounts of heavy metals into water bodies. Mining and agricultural runoff[1] also contribute to water pollution by discharging toxic chemicals into rivers. These activities introduce pollutants that degrade water quality.

River pollution is a significant cause of concern in urban areas as it poses a risk to public health, disrupts ecosystems, and imposes a considerable economic burden. Contaminated rivers can lead to the spread of waterborne diseases, which affects communities that rely on them for livelihood. According to a report by UNICEF and WHO, 1 in 3 people globally cannot access safely managed drinking water [WHO & UNICEF JMP, 2021]. Additionally, harmful substances released into water bodies can kill plant and animal species and alter their habitats. This process destabilises the aquatic ecosystem, leading to long-term environmental consequences. One such major incident was the cyanide spill of 2009 in River Trent, which killed thousands of fish. The economic impact of river pollution is also considerable as local authorities and governments invest heavily in restoring rivers. For instance, the plan for cleaner and more plentiful water, proposed by the Department for Environment, Food & Rural Affairs et al. [2023] in the UK is estimated to cost around £1.6 billion. If pollution management was done diligently, these funds could have been better utilised for essential services, such as the NHS.

1.1 Understanding water pollution

Depending on the source of the pollutants, water pollution can be classified into two categories:

- Point source pollution: Water pollution coming from a single identifiable point source,

[1] Runoff is defined as the draining away of water (and the substances carried in it) from a land's surface.

such as a pipe or a ditch, is called point source pollution. Examples include wastewater discharge from sewage treatment plants and factories, leaking septic systems, and oil spills. Environment agencies in the UK are authorised to regulate point source pollution. They establish limits on what can be discharged directly into the water bodies.

- Non-point source pollution: Non-point source pollution comes from many different sources rather than a single point. It is often caused by the accumulation of pollutants from a large area. Agricultural and stormwater runoff and sediments from construction sites are examples of non-point source pollution. This type of pollution is difficult to regulate since there is no identifiable culprit.

Both point and non-point pollution sources release harmful substances into the water that harm the aquatic environment. Many substances are considered to be water pollutants. Some of the most significant pollutants are listed below:

- *Nutrients* such as nitrogen and phosphorus are common pollutants in municipal wastewater discharges and can lead to rapid biological "ageing" of lakes and streams. Surplus nutrients promote excessive growth of plants and algae in water bodies. When these organisms overgrow and do not get enough sunlight, they deplete the oxygen levels in the water. Furthermore, as large algal blooms decompose, they consume even more oxygen, significantly lowering the water's dissolved oxygen levels and harming aquatic life. This process is known as eutrophication.

- *Sediments and suspended solids* released into a stream due to land cultivation, construction and mining operations may interfere with fish spawning and cause unpleasant odours.

- *Agricultural waste*, including manure and pesticides, is typically high in nutrients like nitrogen and phosphorus, which can significantly degrade surface and groundwater quality.

- *Heavy metals* such as lead, copper, and mercury can bind with organic compounds in water and form detrimental chemicals that can be fatal if ingested.

- *Heated discharges* from industrial effluents or other human activities can raise the water temperature, lowering the solubility of oxygen in the water and reducing the amount of dissolved oxygen available. Heat also increases aquatic organisms' metabolic rate, further depleting oxygen levels.

Most water pollutants are measured in milligrams of the substance per litre of water (mg/L). The Water Framework Directive set out by the European Commission [2024] gives a range of acceptable values of these pollutants for each watercourse in the UK. Along with the specific pollutants, physical parameters like temperature, pH, turbidity (cloudiness), and dissolved oxygen are also used to assess water quality.

1.2 Combined Sewer Overflows

Most of the UK has a combined sewerage system wherein the same pipe carries the wastewater and the rainwater to sewage treatment plants. These sewer pipes can get overloaded during heavy rainfall, exceeding their carrying capacity. In such situations, authorities permit the discharge of untreated waste into water bodies. Combined sewer overflows (CSOs) were developed as safety valves to prevent sewage from backing up during periods of heavy rainfall. CSOs are an example of point source pollution, with 14,326 CSOs present in England. The Environment Agency collaborates with water companies to ensure that they closely monitor and report on their CSO discharge activities.

Despite all preventive measures, CSOs are one of England's leading causes of water pollution. According to a report by the Guardian, untreated effluents were released into England's rivers via combined sewers for more than 1.5 million hours in 2019. Another report by BBC suggested that water firms illegally spill sewage into water bodies on dry days. This practice, known as "dry spelling", is even more dangerous as wastewater is released into the rivers without being diluted. Recently, Yorkshire Water, a company providing water and wastewater services, was fined £1,600,750 for unauthorised sewage discharges into the Bradford Beck (river).

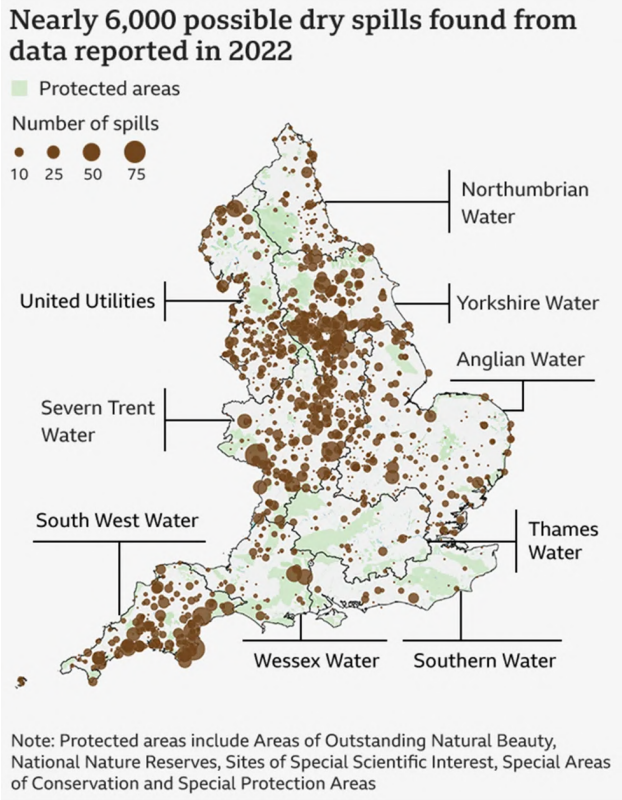


***Figure 1.1*:** Distribution of potential dry spills across England in 2022, with spill frequency indicated by the size of dots. Source: BBC research, Natural England, Ofwat [BBC News et al., 2024].

1.3 Research Background & Objectives

Hydraulics of a river catchment, including flow analysis, transit times, and sediment transport, play a crucial role in understanding the source and distribution of pollutants within the river. Old et al. [2006] highlights urbanisation's impact on Bradford Beck's catchment hydraulics. The study involved continuous flow, turbidity, and specific conductance monitoring at multiple strategic sites along the Beck from June 2000 to June 2001. As a result, replacing natural landscapes with impervious surfaces leads to high concentrations of solutes, increased runoff and elevated sediment transport in the river. Moreover, the urbanised part of the catchment exhibited increased 'flashiness' in both flow and sediment transport, responding rapidly to rainfall events. Data also suggested that the urbanised part of the catchment was the main contributor to river solutes. Similar work done in Goodwin et al. [2003] analyses the sediment transport dynamics in the Bradford Beck catchment for two years, from September 1999 to June 2001. The study uses flow and quality data collected at a high temporal resolution. It was found that urban areas, especially during significant storm events, tend to have higher peak Suspended Sediment Concentration (SSC) than rural areas. Seasonal variation of sediment transport is higher in rural areas than in urban areas, with higher sediment loads during the winters. It was also observed that engineering structures like the Beck Diversion Tunnel and the Esholt Tunnel considerably affect sediment transport by diverting substantial amounts of sediment from the catchment.

Hydrological modelling tools like Soil and Water Assessment Tool (SWAT) [Arnold et al., 1998] and MIKE 11 [DHI, 2007] are widely used to simulate river discharge and pollution loads in water bodies. Tran et al. [2017] investigates nitrogen pollution in the Cau River Basin, one of Vietnam's most polluted river basins. SWAT was applied to simulate streamflow and nitrogen loads from various sources like agriculture, households, industry, craft villages, and livestock to identify the major contributors to nitrogen pollution in the basin. After analysing three scenarios with different combinations of pollution sources, it was established that treating small-scale industries, craft villages, and livestock as non-point sources provided the best simulation of nitrogen loads. This scenario also showed a strong correlation between rainfall and nitrogen load, with the wet season contributing significantly more to nitrogen pollution than the dry season. Researchers also identified cultivation and livestock as the most significant contributors to nitrogen pollution, followed by small-scale industries and craft villages.

The Concentration-Discharge (C-Q) relationship is also crucial in identifying the source of the pollutants. C-Q relationships are used as 'hydrochemical tracers' to determine the variability in solute export across different time scales. Van Emmerik et al. [2022] focuses on understanding the role of hydrology in the movement of plastic debris within the Rhine-Meuse delta in the Netherlands over one year. It was found that plastic transport in rivers is significantly influenced by hydrological events, like peak discharges and floods, which can increase plastic transport by up to six times. Researchers also identified spatial variations in plastic transport due to urban areas, tidal dynamics, and network complexity. Hashemi et al. [2020] uses Principal Component Analysis (PCA) and C-Q relationship to analyse nutrient export behaviours in mini-catchments

across Denmark, Finland and Sweden using 8-year data series. It classifies the C-Q relationship into nine types based on the export regime - enrichment (an increase of C with Q), constant (no significant relationship between C and Q) and dilution (decrease of C with Q), and hysteresis (time lag) - clockwise, no hysteresis, and anticlockwise. Findings revealed that nutrient export behaviour in these streams is dominantly controlled by air temperature and land use and, to a lesser extent, by their climate.

Poor identification and quantification of pollution sources are the leading cause of ineffective pollution management. This study aims to understand the catchment[2] hydraulics of a small river system in Bradford, UK, to provide insights into the source and dynamics of the river pollution events. The specific objectives are:

- Obtain and organise flow, quality, and rainfall data for the period of study;
- Estimate the catchment area for each tributary and each flow gauge site.
- Calculate how flow accumulates along the river in dry and wet weather and relate this to the increases in the catchment area.
- Identify notable water quality events and use these to calculate transit times.

The datasets for this study are provided by Yorkshire Water's UPM (Urban Pollution Management) study. These datasets are then processed and visualised using exploratory methods. Catchment area estimation, a crucial part of the study, is carried out using GIS (Geographic Information System), an advanced technique that combines maps with databases to help users create, manage and analyse geographic information. Flow accumulation is then calculated by finding the cumulative flow at each site and is presented with appropriate visualisations. For identifying the water quality status of the river, we will use Total Ammonia Nitrogen (TAN)[3] as the primary pollutant. Transit times are then calculated by finding the time difference between TAN peaks at each site.

[2]River catchment is the area of land where all surface water drains to a single river or stream.

[3]TAN (total ammonia nitrogen) is the total amount of nitrogen in the forms of NH3 (unionised) and NH4+(ionised) in the water.

Chapter 2

Data preparation and EDA

2.1 Study area

The study will be conducted in the small (60km^2), heavily urbanised catchment of the Bradford Beck, located in Bradford, West Yorkshire. Bradford Beck is a small river system of around 11km that flows through the city. It starts as a collection of small tributaries in the west, which flow eastwards toward the city centre, forming the Bradford Beck. The river takes a northern turn from the city centre towards Shipley, where it has its confluence with river Aire. Most of the beck is culverted (inside tunnels) near the city and has limited accessibility.

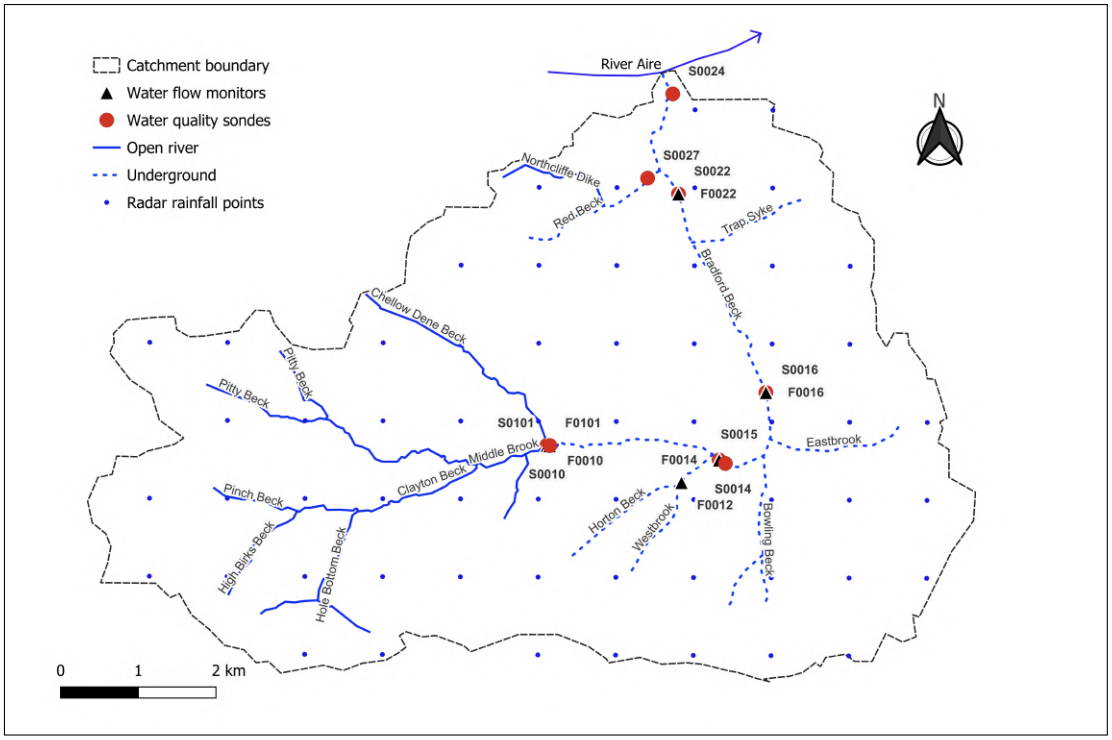


***Figure 2.1*:** Schematic map of the Bradford Beck and locations of flow, quality and rainfall monitoring sites along the beck.

During the Industrial Revolution, Bradford Beck was used as a sewer to carry away industrial and domestic waste, which resulted in substantial water pollution. Since then, there have been many attempts to improve water quality in the beck, but significant changes have yet to be observed. Today, CSO discharges and surface water runoff are the primary sources of pollution in the beck. There are 50 combined sewer overflows installed along the beck. As a result, Bradford Beck is now classified as ”poor ecological quality” under the Water Framework Directive.

***Figure 2.2*:** CSO located approximately 2 km upstream from the Shipley station. Image procured during the field trip organised to Bradford beck on June 6th

2.1.1 FoBB and ART

This project is in collaboration with the Friends of Bradford's Beck (FoBB) and Aire Rivers Trust. FoBB (founded in 2012) are a group of Bradford-based residents and interested ecologists who work towards the restoration of the Bradford Beck. They actively organise guided walks, litter picking sessions, funding campaigns, and training courses to improve the beck. ART is a charity dedicated to improving the quality of rivers that flow through the Aire Valley. FoBB is affiliated with ART.

2.2 Data understanding

In 2020, Yorkshire Water installed some flow and quality monitors around the catchment, recording data at a high temporal resolution every 15 minutes. The monitors have been strategically placed along the beck (see appendix A), and data has been recorded for 381 days, from 23-09-2019 to 07-10-2020. Radar rainfall data for Bradford, gridded at one sq. km, is also obtained for the same period. All monitoring points' locations are shown in Figure 2.1.

***Figure 2.3*:** Flow and quality monitoring device used for data collection.

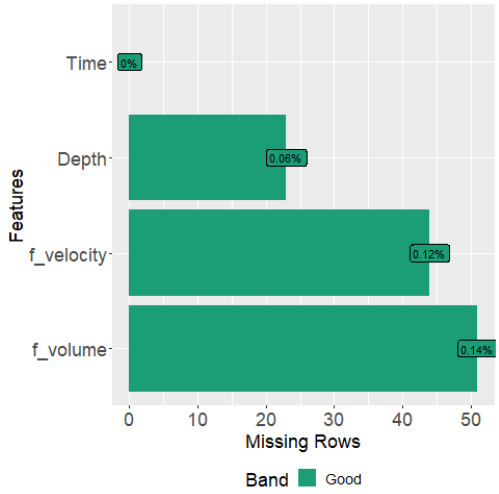
2.2.1 Flow data

There are six flow monitoring sites recording data at every 15-minute interval. The data dictionary for the flow datasets is defined below.

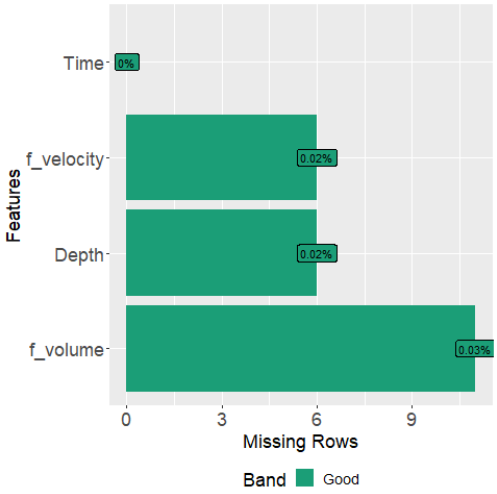
***Table 2.1*:** Data dictionary for flow datasets. Depth and velocity are the mean values recorded at different points on the site. Flow volume is calculated as cross-sectional area$\times$velocity.

Variable	**Data type**	**Description**
Time	Date Time	The date and time of recording the flow observations (at 15 min intervals).
Depth	Numeric	The depth of the river in mm.
f_velocity	Numeric	The flow velocity of the river in m/s.
f_volume	Numeric	The flow volume/discharge of the river in litres/sec.

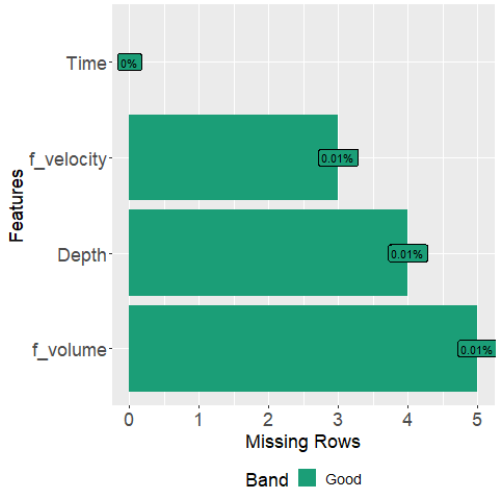
Data quality checks confirmed that there are missing values present in the data. The proportion of missing values in the datasets is plotted using the `DataExplorer` package in R.



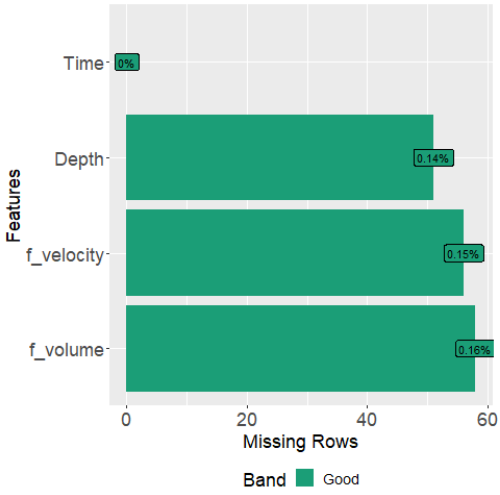
(a) Site F0010 (Total missing: 0.32%).



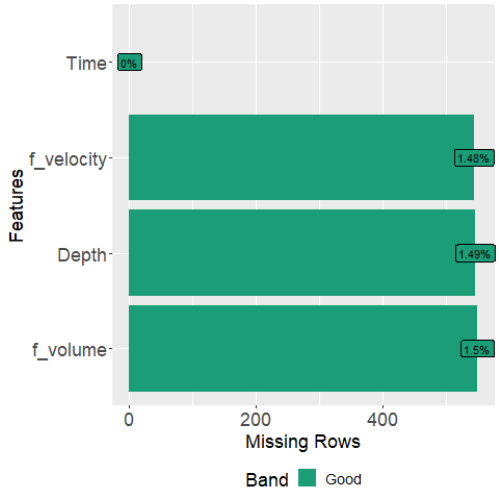
(b) Site F0012 (Total missing: 0.06%).



(c) Site F0014 (Total missing: 0.03%).



(d) Site F0016 (Total missing: 0.45%).



(e) Site F0022 (Total missing: 4.47%).

(f) Site F0101 (Total missing: 4.52%).

***Figure 2.4*:** Plots showing the missing values in each flow dataset.

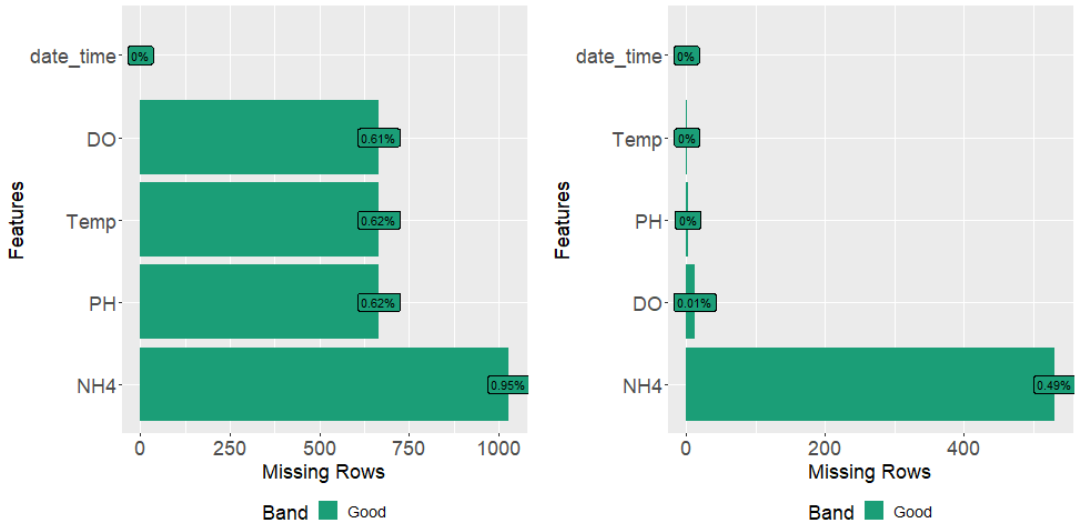
2.2.2 Quality data

There are eight quality monitoring sites gathering data at every 15-minute interval. The data dictionary for the quality datasets is defined below.

***Table 2.2*:** Data dictionary for quality datasets.

Variable	Data type	Description
date_time	Date Time	The date and time of recording the quality observations (at 15 min intervals).
DO	Numeric	Amount of dissolved oxygen in the river in mg/L.
NH4	Numeric	Total ammonia nitrogen (NH_3 + NH_4^+) in the river in mg/L.
PH	Numeric	pH value of the water in the river.
Temp	Numeric	Temperature of the river water in °C.

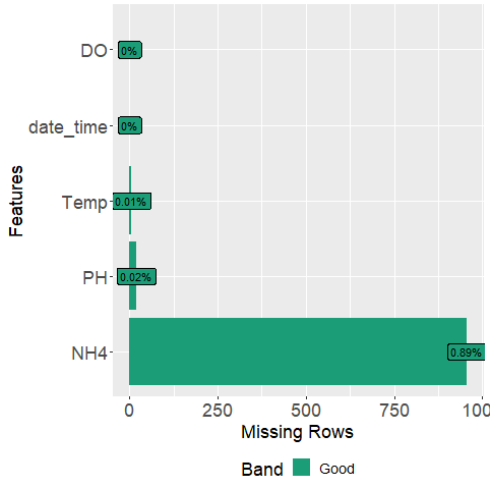
There are missing values present in the quality datasets. The proportion of missing values in each dataset is shown below.



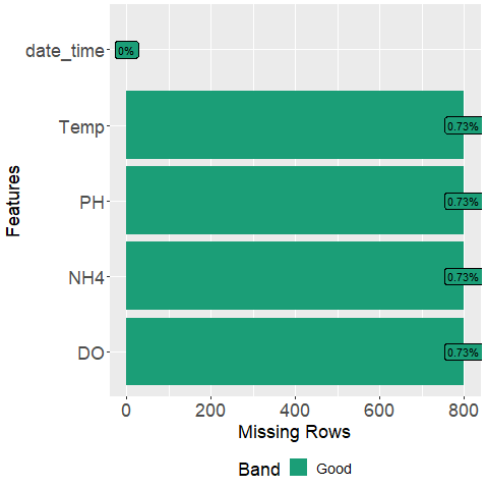
(a) Site S0010 (Total missing: 2.79%).

(b) Site S0014 (Total missing: 0.50%).

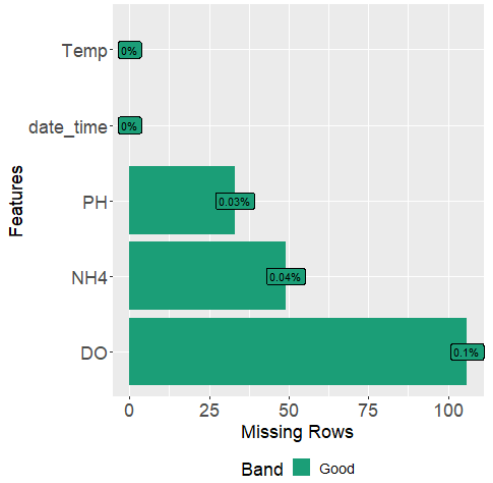
***Figure 2.5*:** Plots showing the missing values in quality datasets for sites S0010 and S0014.



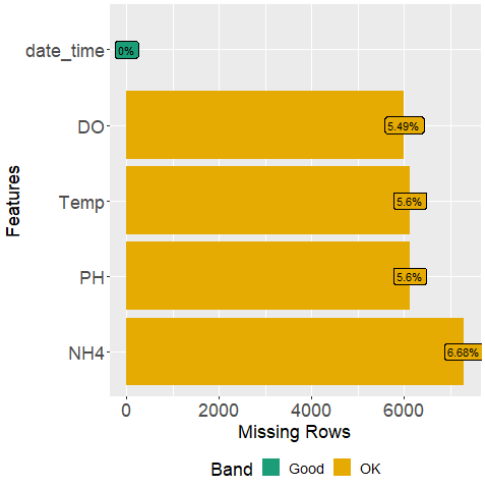
(a) Site S0015 (Total missing: 0.91%).



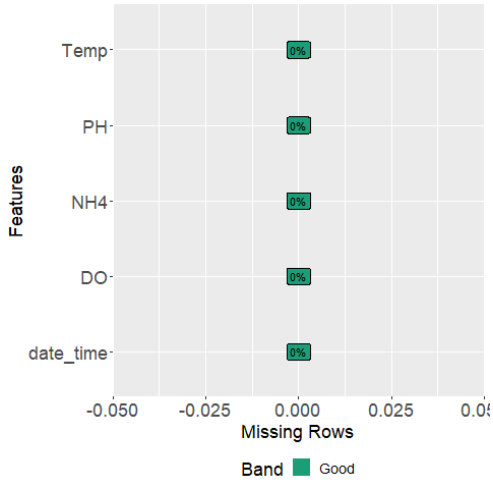
(b) Site S0016 (Total missing: 2.92%).



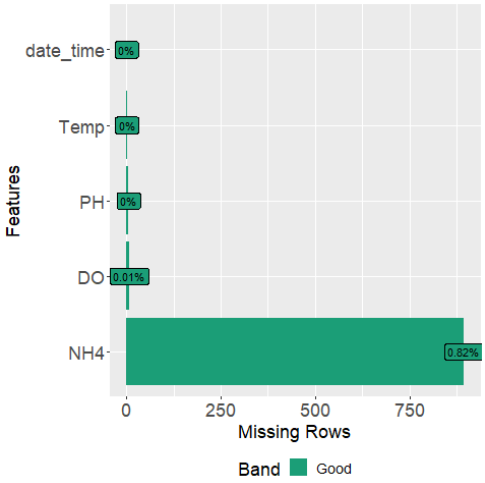
(c) Site S0022 (Total missing: 0.17%).



(d) Site S0024 (Total missing: 23.38%).



(e) Site S0027 (Total missing: 0%).



(f) Site S0101 (Total missing: 0.83%).

***Figure 2.6*:** Plots showing the missing values in quality datasets for sites S0015, S0016, S0022, S0024, S0027, and S0101.

2.2.3 Imputation

Since removing the missing values would result in a biased dataset, missing values must be imputed using an appropriate imputation technique. Many studies, including Anil Jadhav and Ramanathan [2019] and Gautam and Latifi [2023], indicate that K Nearest Neighbor (KNN) consistently outperforms other imputation algorithms in the case of simple numeric data. Hence, we will be using KNN imputer to fill in the missing values in all the datasets. For detailed working of the KNN imputer, see Appendix B.

Choice of K

The value of k depends on the input data characteristics and is a tradeoff between bias (high value of k) and variance (low value of k). We will tune the value of k for each dataset using the elbow method. The following procedure is carried out for each dataset.

Step-1: Remove all rows containing NA values and apply standard scaling[1] on the dataset.

Step-2: Randomly[2] substitute 5% of the entries as missing values in the scaled dataset.

Step-3: Fit KNN imputer with k=1 to 10 on the missing data.

Step-4: Calculate the Mean Squared Error of the imputed values using the scaled dataset for all k.

Step-5: Plot MSE vs k and find the elbow point of the plot.

Step-6: Use the k value obtained to fit the KNN imputer and inverse the scaling to get the final imputed dataset.

The MSE vs k plot for flow site 5 (F0022) is shown below.

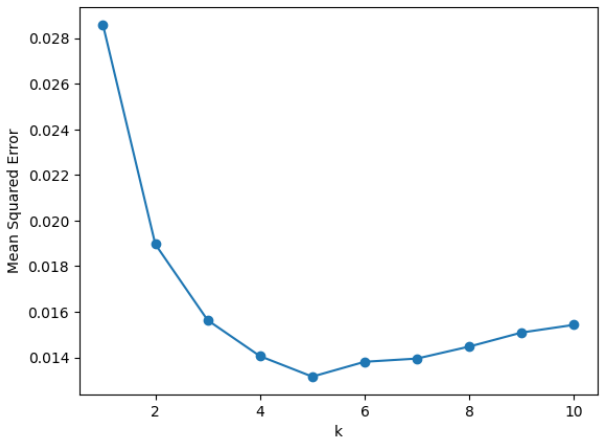


***Figure 2.7*:** Mean squared error vs k plot for F0022 site data. It can be seen that the elbow point is located at k=5.

[1] Standardises the data to have a mean of 0 and a standard deviation of 1, $z = \frac{x-\mu}{\sigma}$

[2] The data is assumed to be missing completely at random (MCAR).

The optimal value of k for all the flow and quality datasets is obtained using this method. For all plots and values, see Appendix B. KNN imputer is fit on all datasets using the optimal k values.

2.2.4 Rainfall data

Rainfall data provided by Yorkshire Water had a low spatial coverage. Hence, we used radar rainfall data (recorded daily) gridded at $1km^2$ obtained from the CEDA (Centre for Environmental Data Analysis) archive [Met Office et al., 2021], covering the period from September 2019 to October 2020. The original dataset was in netCDF(network Common Data Form) format, then converted to CSV (Comma Separated Values) format for analysis. The data dictionary for the final rainfall dataset is defined below.

***Table 2.3*:** Data dictionary for rainfall dataset.

Variable	Data type	Description
time	Date	Date of recording the rainfall observation (daily recordings).
latitude	Numeric	Latitude of the recording site in degrees.
longitude	Numeric	Longitude of the recording site in degrees.
rainfall	Numeric	Depth of rainfall recorded in mm.

No data quality issues were found, and the dataset was complete.

2.3 Exploratory Data Analysis

The following plots were produced to understand the data better and draw some preliminary insights from it.

2.3.1 Average daily flow vs rainfall for all flow monitoring sites.

Figure 2.8 shows that flow positively correlates with rainfall depth. The data points are mostly clustered towards the origin, indicating that most days experience little/no rainfall with low flow values. There are also outliers in the data, which indicate 'flashy' flows[3] and might give information about the land use at each flow site.

[3]Replacement of natural surfaces with hard, impervious surfaces causes quick runoff directly into rivers or drains. This runoff rapidly increases and decreases river flow, referred to as 'flashy' flow.

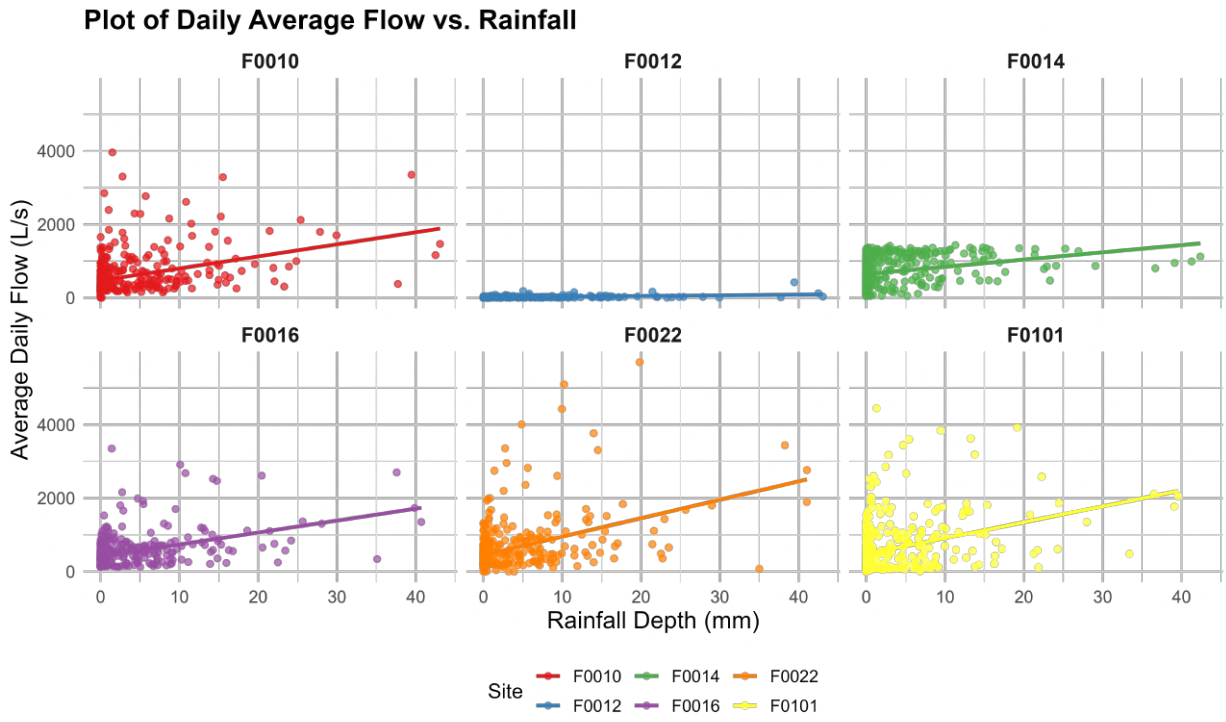


***Figure 2.8*:** The daily average flow for all flow sites is plotted against the rainfall depth. *Rainfall depth* is the average rainfall recorded in the four nearest radar points from the respective flow sites. A simple linear fit is also done on the data.

2.3.2 Total ammonia nitrogen profile

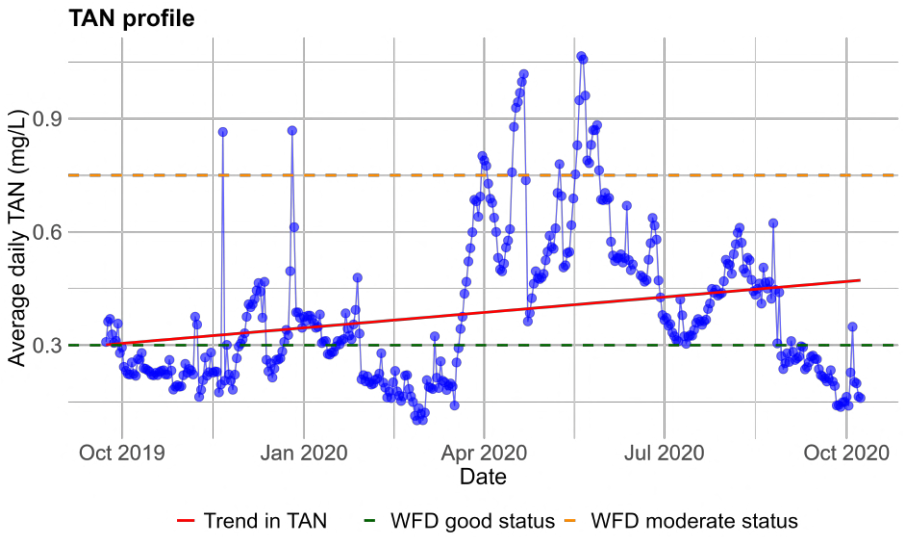


***Figure 2.9*:** Daily average TAN values against time for an upstream quality monitoring site (S0010). The plot also contains threshold values for TAN as defined by the Water Framework Directive.

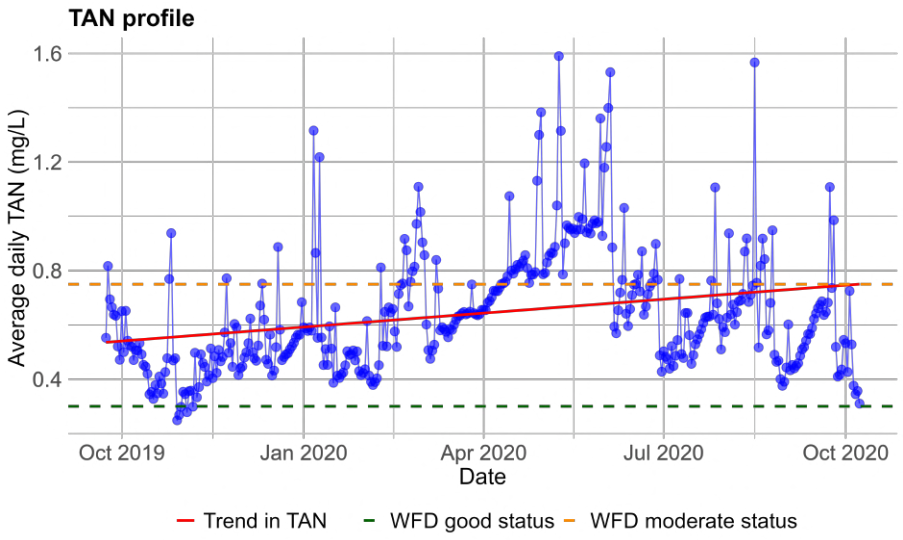


***Figure 2.10*:** Daily average TAN values against time for S0016 quality monitoring site located in the middle of the catchment. The plot also contains threshold values for TAN as defined by the Water Framework Directive.

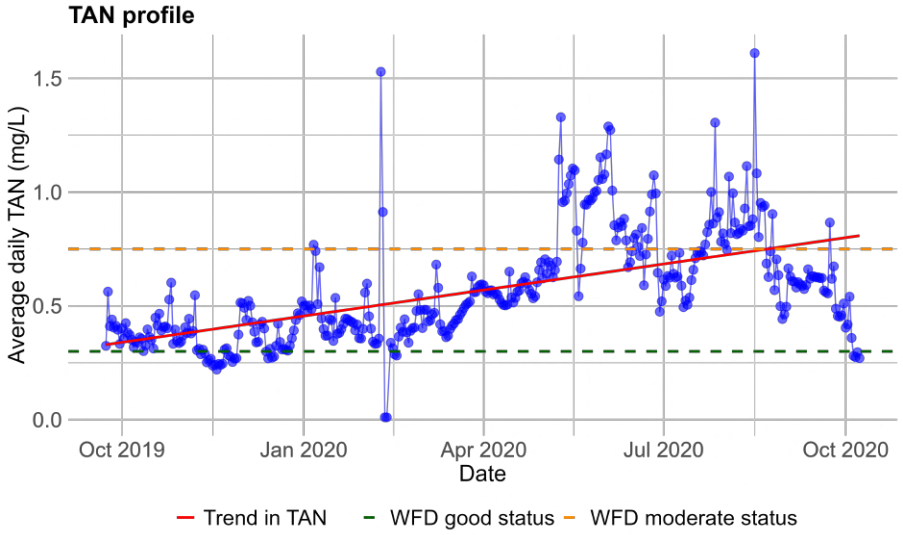
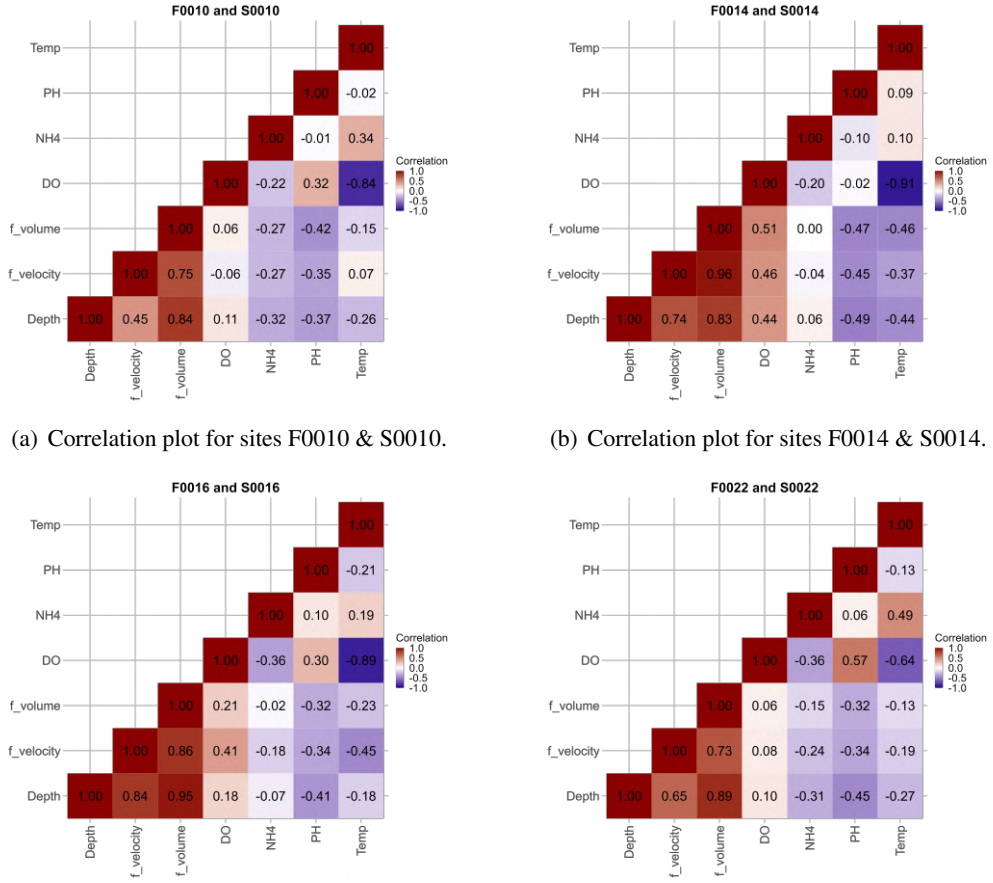


***Figure 2.11*:** Daily average TAN values against time for a downstream quality monitoring site (S0022). The plot also contains threshold values for TAN as defined by the Water Framework Directive.

Figures 2.9, 2.10, and 2.11 indicate that the global trend in TAN increases over time. It also reveals that the 'good' threshold (0.3 mg/L) for water quality is violated almost constantly during the period.

2.3.3 Correlation heatmap



(a) Correlation plot for sites F0010 & S0010.

(b) Correlation plot for sites F0014 & S0014.

(c) Correlation plot for sites F0016 & S0016.

(d) Correlation plot for sites F0022 & S0022.

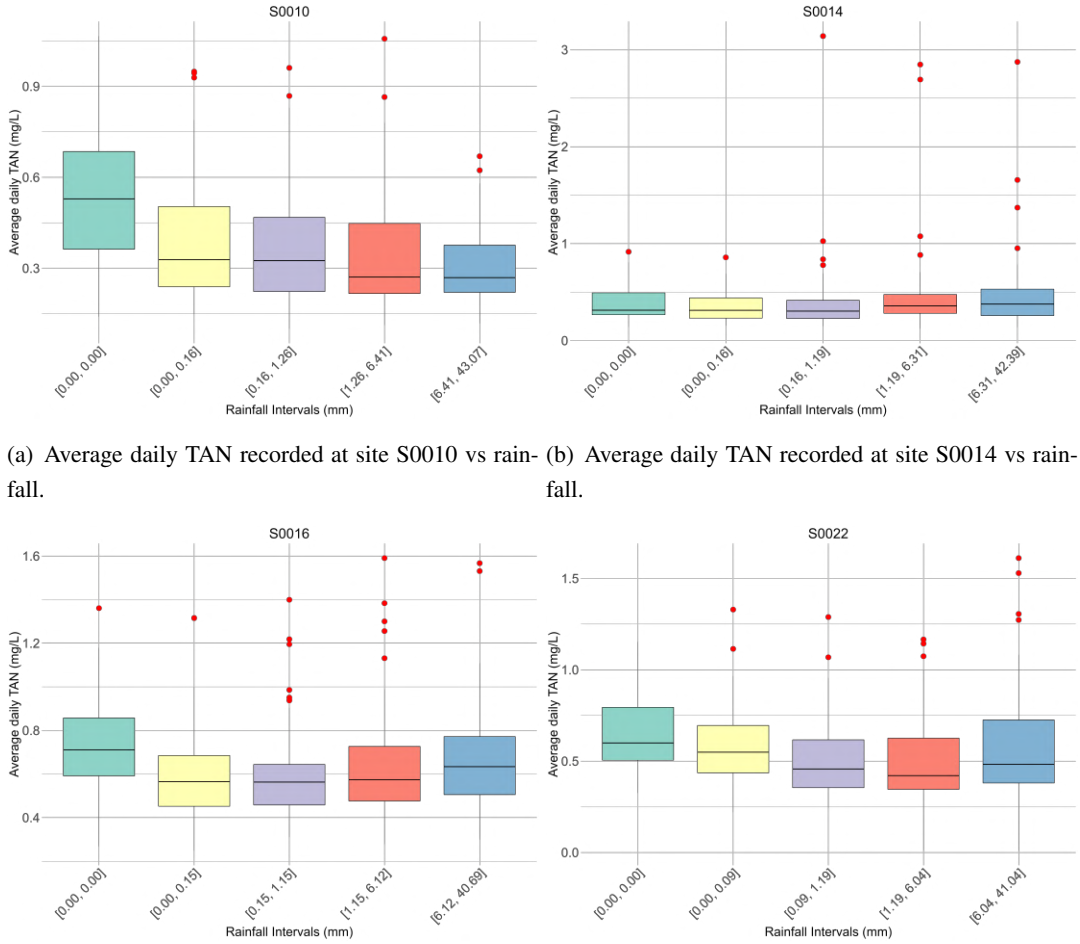
***Figure 2.12*:** Heatmap showing the correlation between variables at common flow and quality monitoring sites.

The correlation plots in Figure 2.12 confirm the sanity of the data. It can be seen that Dissolved Oxygen is highly negatively correlated with temperature, which is expected since higher temperatures lower the solubility of oxygen in the water. Temperature negatively correlates with flow since higher flows transfer more heat and cool the liquid. Interestingly, pH is also negatively correlated with flow volume, which can be due to the fact that during high flow events (such as heavy rainfall), increased runoff carries more pollutants into the river, making the water more acidic.

2.3.4 Average daily TAN vs rainfall

Boxplots in Figure 2.13 show that the average TAN value is highest during dry weather at all monitoring stations, indicating the spilling of CSOs even in dry weather. The TAN value again becomes high for S0016 and S0022 at high rainfall but goes down for S0010. This might be

because the catchment area of S0010 is less urbanised, leading to lower pollution by surface runoff during rainfall events. Outliers are also present in the data and are mainly observed during wet weather.



(a) Average daily TAN recorded at site S0010 vs rainfall.

(b) Average daily TAN recorded at site S0014 vs rainfall.

(c) Average daily TAN recorded at site S0016 vs rainfall.

(d) Average daily TAN recorded at site S0022 vs rainfall.

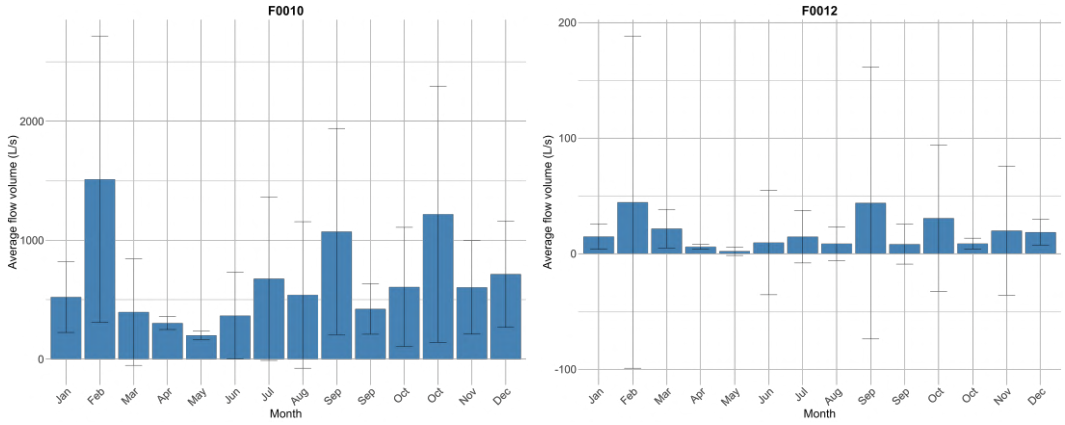
***Figure 2.13*:** Boxplot showing the average daily Total Ammonia Nitrogen (TAN) recorded at four quality monitoring sites along with the rainfall intervals.

2.3.5 Average monthly flow volume

Figure 2.14 shows the monthly average flow volume at all flow sites. The flow increases downstream and is the highest at F0014. After that, it decreases a bit[4]. This could be due to the diversion tunnel being created near the city centre to divert the river's flow. The whiskers in the plot show the variation in the flow values during each month. It can be seen that for some months, the whiskers have become negative. Negative values imply that the standard deviation

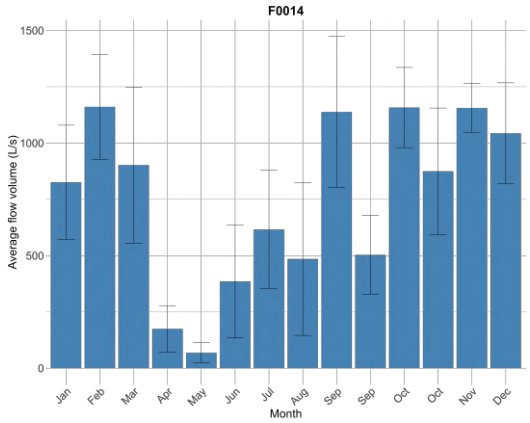
[4]The flow of a river generally increases as we go downstream unless there are losses or diversions. Discharge increases downstream because of additional water from tributaries.

of the flow for that month is greater than the mean flow, indicating highly variable, 'flashy' flows.

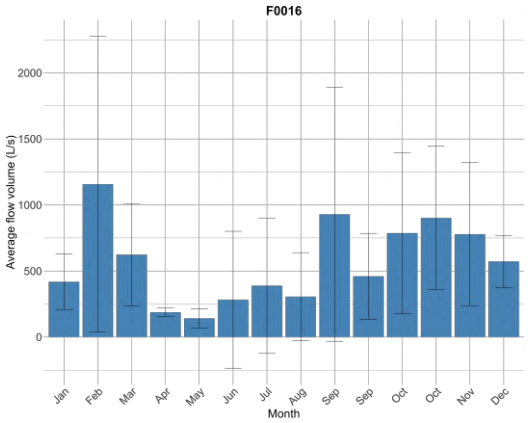


(a) Site F0010.

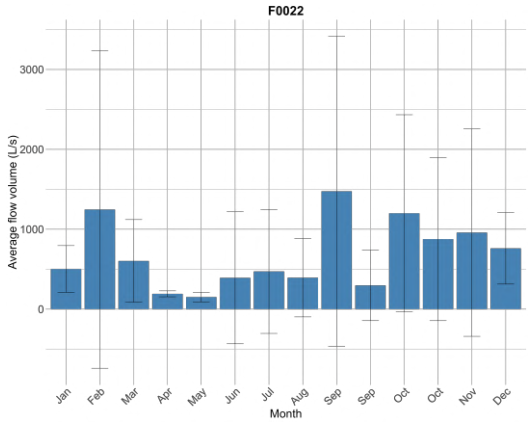
(b) Site F0012.



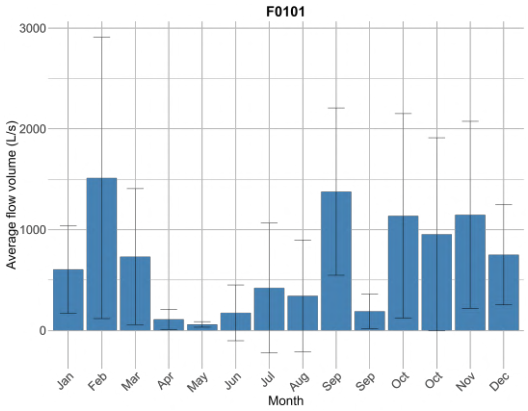
(c) Site F0014.



(d) Site F0016.



(e) Site F0022.



(f) Site F0101.

***Figure 2.14*:** Plots showing the average monthly flow at all flow monitoring sites. The whiskers in the plots show the variance in the flow values (mean + S.D., mean - S.D).

Chapter 3

Catchment area estimation

A river's catchment area, also known as a drainage basin, is the land from which water collects and flows into the river. Catchment area estimation is crucial in water quality management by providing insights into pollutants' sources and transport mechanisms. Models like Storm Water Management Model (SWMM) [U.S. Environmental Protection Agency, 2012], used to simulate pollutant loads and source area contributions, and CatStream model [Hossain and Imteaz, 2013], which integrates pollutant processes with hydrological dynamics, use catchment area to enhance their modelling approach. Catchment area characteristics like land use, land cover, and topography directly influence how water and pollutants travel through the landscape.

In this section, we will be estimating the catchment area of:

- Bradford beck and each of its tributaries to better understand the overall channel hydraulics.
- Each flow monitoring gauge to calculate the flow accumulation in the next section.

3.1 Methodology

The complete workflow for delineating the catchment area of a stream using Digital Elevation Model data is below.

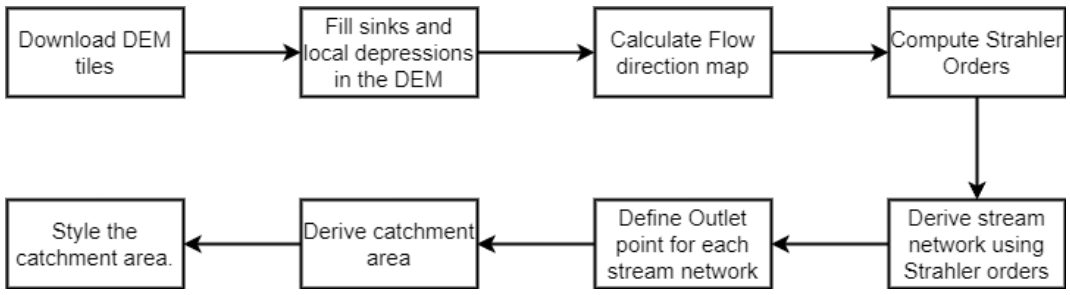


***Figure 3.1*:** Complete workflow for catchment area delineation using Digital Elevation Model (DEM) data.

3.1.1 Data sources and GIS tools:

Digital Elevation model data is used in this section to estimate the catchment area. A digital elevation model (DEM) is a 3-D representation of the bare Earth's surface, excluding trees, buildings or any surface objects. A DEM is generated by collecting elevation measurements across a surface and storing them in a raster dataset, which consists of a regular grid of pixels containing elevation values. Different techniques can be used to create a DEM, with the most common ones being LiDAR (Light Detection and Ranging) and Stereo Photogrammetry (using overlapping photos to create a 3D model). We obtain the DEM data for Bradford (at 30m resolution) using satellite data collected from NASA's Shuttle Radar Topography Mission (SRTM) [NASA Earth Science Data Systems (ESDS), 2024].

Geographic Information System (GIS) techniques will effectively delineate the catchment area and visualise the spatial data. All analyses are conducted in QGIS, a free and open-source software for analysing geospatial information. It supports vector and raster data formats and is built on top of Python, allowing users to integrate QGIS with other Python-based tools. We will use the PCRaster tools plugin [Karssenberg et al., 2010] in QGIS, which has extensive functionality for hydrological modelling.

3.1.2 Preprocessing DEM and computing flow directions:

After loading the DEM, the next step is to fill in the local sinks in the data. Sinks are pits or depressions in a DEM that obstruct the natural drainage of water downstream, caused by an error in recording or the presence of natural lakes and ponds. The presence of sinks in the data may result in an erroneous flow-direction raster.

Next, we calculate the flow direction raster from the filled DEM. To achieve this, we use the `lddcreate` function of the PCRaster tools plugin which integrates sink removal and flow direction calculation. `lddcreate` uses the D8 algorithm [O'Callaghan and Mark, 1984] to compute flow direction by evaluating the slope of each of the 'eight' neighbouring cells from a given cell and assigning the flow direction arrow to the neighbour with the steepest downslope.

Given a Digital Elevation Model with the current cell located at (i, j) and its neighbouring cell at (i', j'), the formula for computing the slope from the current cell to the neighbouring cell is:

$$\text{Slope} = \frac{\text{Elevation}_{ij} - \text{Elevation}_{i'j'}}{d} \tag{3.1}$$

where:

- Elevation_{ij} is the elevation of the current cell.
- $\text{Elevation}_{i'j'}$ is the elevation of the neighboring cell.
- d is the distance between the current cell and the neighbouring cell.

In a regular grid with square cells of side length s, the distance d between a cell and its immediate (N, E, S, W) neighbour is s, and the distance to diagonal neighbours (NE, NW, SE, SW) is:

$$d = \sqrt{2} \cdot s \tag{3.2}$$

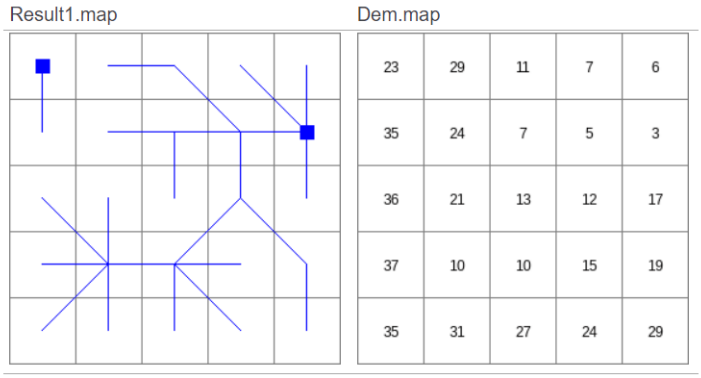


***Figure 3.2*:** Flow direction map (left) for a given DEM (right) derived using the D8 algorithm [PCRaster, n.d.] in PCRaster.

Each cell in the flow direction raster is assigned a value from 1-9, with 5 being the centre cell, 1 being the southwest direction, 2 being the south, 3 being the southeast direction and so on. The directional encoding used by `lddcreate` function is shown below.

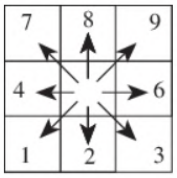


***Figure 3.3*:** Directional encoding of `lddcreate`. A value of 5 is assigned to the centre cell.

3.1.3 Deriving stream channels using Strahler Orders (for river catchment only):

For river catchment delineation, we need to obtain the channel of rivers from the stream network using the flow direction map. Stream network classification was initially developed by Horton and later modified by Strahler [Strahler, 1964]. `streamorder` function of PCRaster assigns an order to each cell in the flow direction map. It designates an order of 1 to the most minor channels, which are the cells to which no upstream cells are connected. When two channels of order 1 join, a channel of order two is formed downstream. In general, when two cells of order i join, a channel of order i+1 is formed downstream.

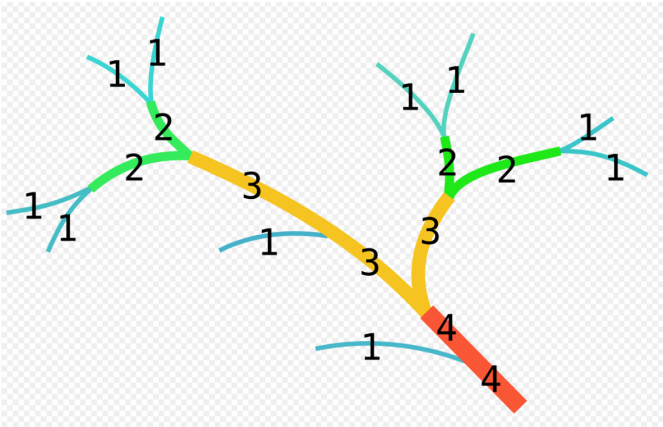


***Figure 3.4*:** Strahler Orders assignment example.

All streams with Strahler Order greater than six are filtered out (threshold found through trial and error) and used as a representation of the Bradford beck and its tributaries.

3.1.4 Defining outlet points and delineating catchment:

Outlet point is defined as the lowest (farthest downstream) point of a river/stream. We define the coordinates of the outlet point for each tributary (and gauge) by approximating its location on the map. To ensure that the outlet point is actually located on the channel raster, we use the `snap pour point` functionality in QGIS. This adjusts the outlet point by snapping it to the location with the highest flow accumulation[1] in the channel within a specified distance.

Using the snapped outlet points and the flow direction raster, we delineate the catchment areas using the `subcatchment` function available in PCRaster tools. Finally, we apply the desired styling to the map.

3.2 Results

3.2.1 River catchment area

The final catchment map and catchment areas derived for Bradford Beck and its tributaries are shown below.

[1]Flow Accumulation in GIS is calculated as the number of cells (pixels) that flow into a given cell based on the flow direction raster.

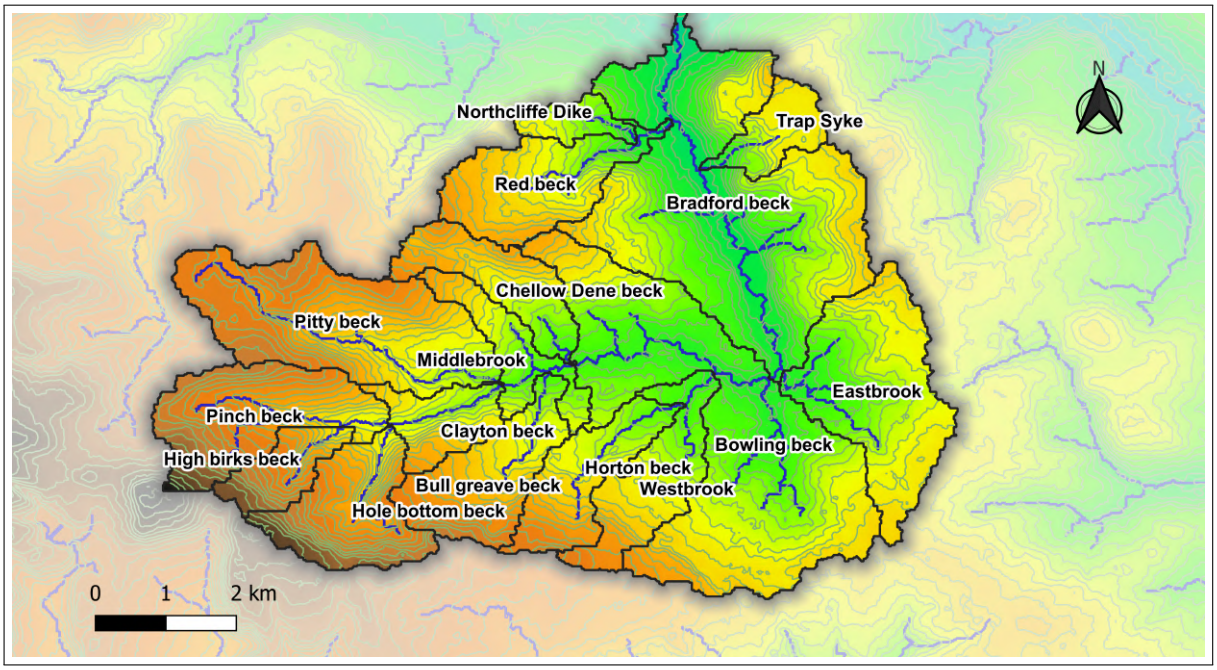


***Figure 3.5*:** Map showing the delineated catchment area for Bradford Beck and each of its tributaries. Contour interval: 10m

***Table 3.1*:** Catchment area of Bradford Beck and its tributaries.

Name	**Area (km^2)**
Bradford Beck	11.17
Bowling Beck	10.82
Pitty Beck	5.97
Eastbrook	4.84
Pinch Beck	3.64
Red Beck	3.01
Hole Bottom Beck	2.9
Bull Greave Beck	2.84
Westbrook	2.51
Horton Beck	2.29
Chellow Dene Beck	2.13
Middle Brook	1.58
Clayton Beck	1.47
High Birks Beck	1.28
Trap Syke	1.26
Northcliffe Dike	0.72

3.2.2 Flow gauge catchment

The final catchment map and catchment areas for each flow gauge[2] are shown below.

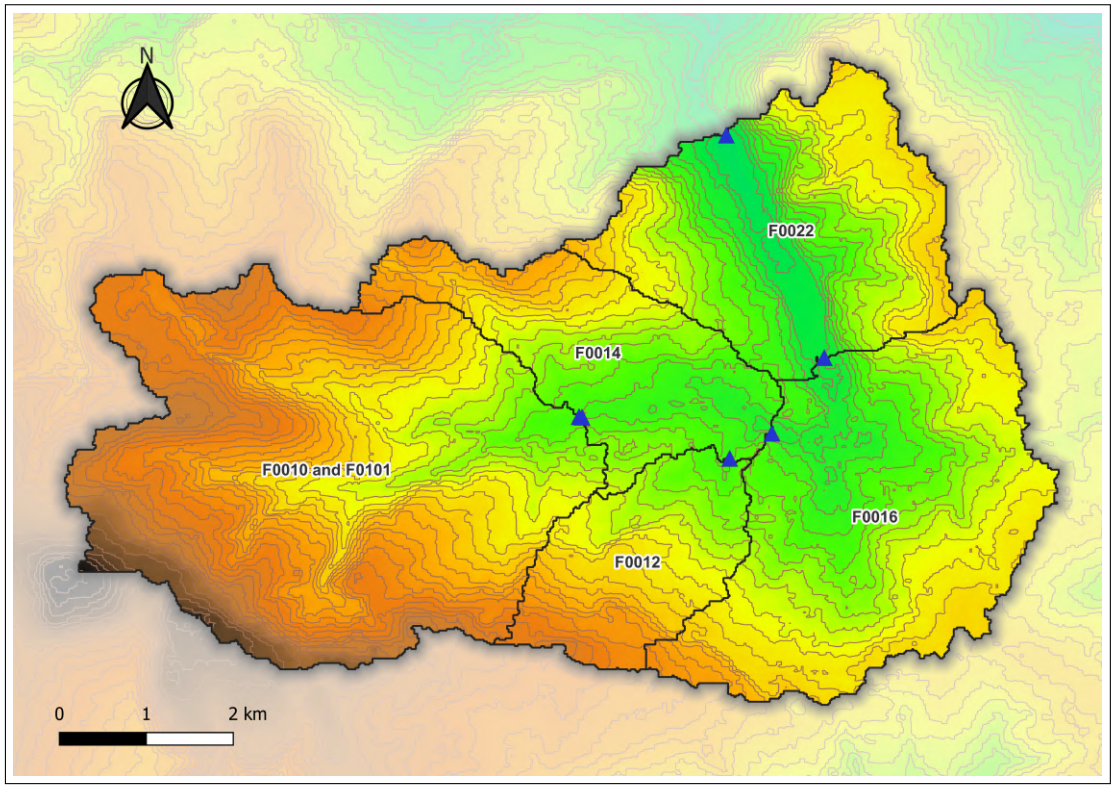


***Figure 3.6*:** Map showing the catchment area for all flow gauges.

***Table 3.2*:** Catchment area for each flow monitoring site. F0010 and F0101 are very close together and have the same catchment area. Contour interval: 10m

Flow Site	**Catchment area (in km^2)**
F0010	19.67
F0012	4.54
F0014	6.15
F0016	12.46
F0022	9.50
F0101	19.67

[2]The catchment area of each flow gauge is defined as the geographic area from which all surface water flows into that specific flow gauge site.

Chapter 4

Flow Accumulation analysis

Flow accumulation refers to the total volume of water that collects downstream in a river per unit of time, originating from all upstream sites and tributaries. It aids significantly in investigating pollution by providing insights into the transport and distribution of pollutants along the river. It can be integrated with water quality models like Soil and Water Assessment Tool (SWAT) [Arnold et al., 1998] to predict pollutant concentrations and identify pollution hotspots in the catchment. Flow accumulation also plays a pivotal role in assessing the Water Environmental Capacity (WEC) of a river [Wang et al., 2023]. WEC provides a quantitative framework for assessing the maximum allowable pollutant load that a water body can accommodate while meeting established water quality standards.

Flow accumulation combined with catchment area and rainfall data can also provide insights into the runoff rate [1]. By relating each gauge site's catchment area to the site's flow accumulation, it can be inferred whether the flow results from a uniform or non-uniform runoff rate. This relation can further enhance our understanding of the catchment hydraulics and also aid in providing flood mitigation strategies.

4.1 Flow duration curves:

The flow Duration Curve represents the relationship between the flow magnitude at a site and the proportion of times that flow was equalled or exceeded over a specific period. It gives information about the variability and frequency of flow rates in the river. The area under the flow duration curve gives the average daily flow value. It is widely used in water resource management and hydraulic structure design.

Flow duration curves for all flow sites are plotted using the `ggplot2` package in R. To enhance interpretability, each flow observation is divided by the maximum flow value observed during the period (15,403.93 L/s), scaling the observations to values between 0 and 1.

[1] Runoff rate is the proportion of rainfall falling that flows into the stream.

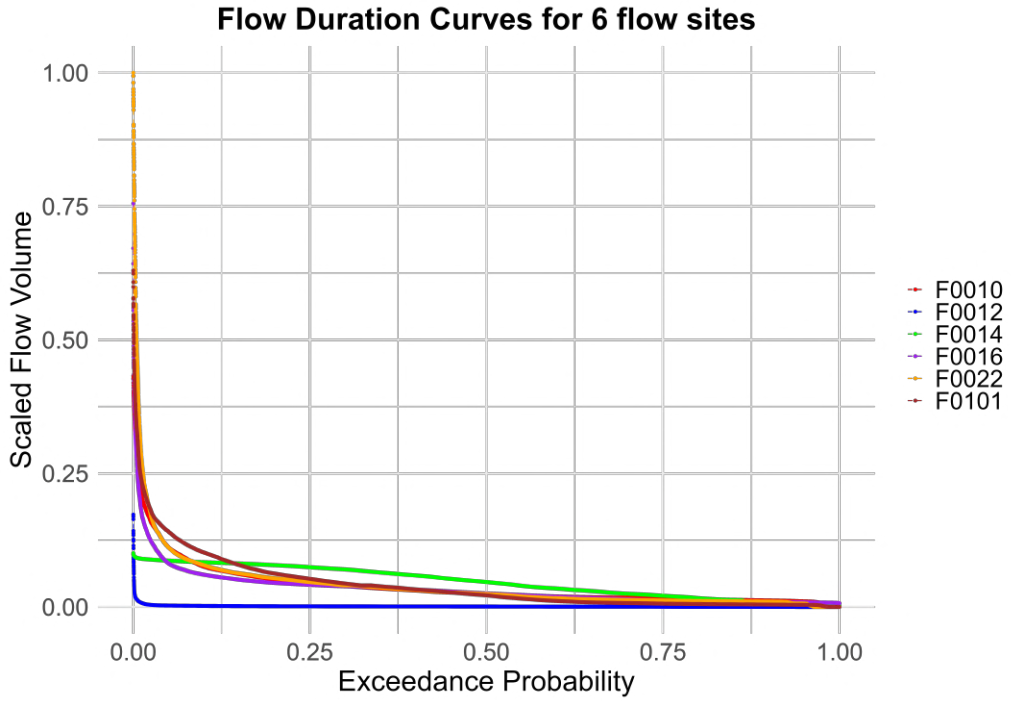


***Figure 4.1*:** Flow duration curve for all 6 flow sites. The y-axis represent the value of scaled flow and the x-axis represents the probability of that flow being exceeded at the particular site.

All flow duration curves have a steep initial slope, indicating significant high-flow events that occur infrequently. As the exceedance probability increases, the curves flatten, indicating that the flow is much lower and relatively more consistent most of the time. F0012 observes a significantly low flow across all exceedance probabilities. F0014 maintains a high flow value at most exceedence probabilities implying that it may experience consistently higher flow volumes. All other sites experience similar flow patterns.

4.2 Methodology

4.2.1 Cumulative flow calculation:

Due to the 'flashy' nature of flows in urban regions, we use cumulative flow to determine the flow accumulation at a site and avoid any anomalies. Cumulative flow at a site is calculated by adding together flows from all upstream sites, including tributaries. It provides a more accurate representation of each site's contribution to the total flow.

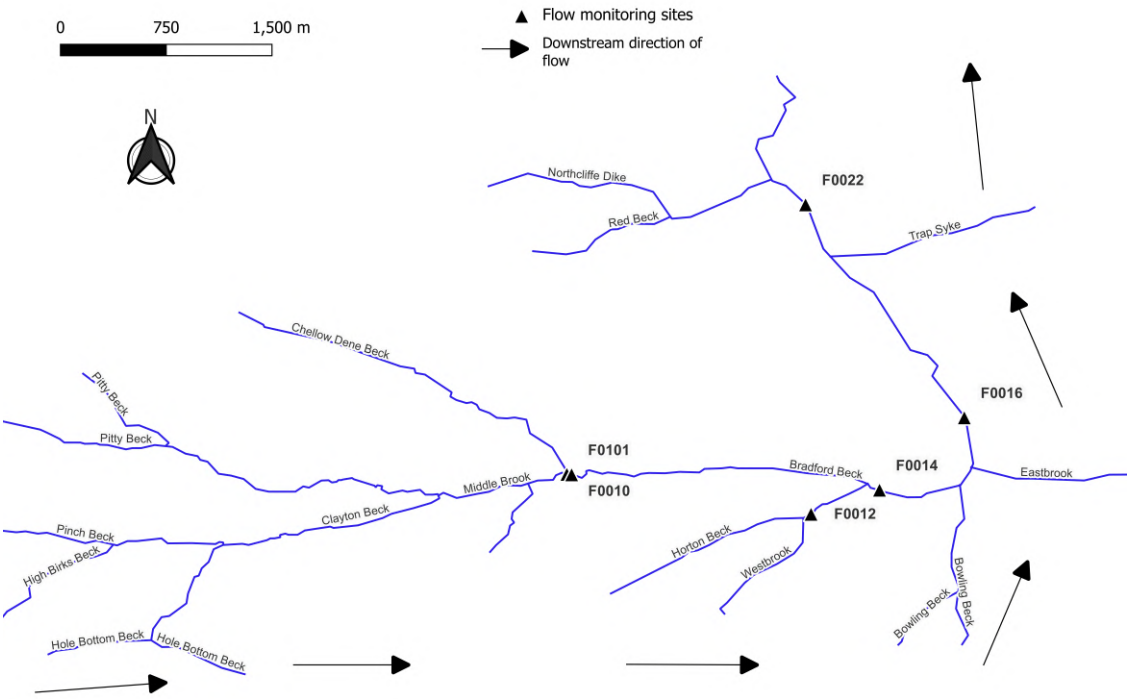


***Figure 4.2*:** Plot showing the flow monitoring sites along with river flow direction.

Cumulative flow for all sites is calculated as follows:

1. Cumulative flow for F0010 = Flow for F0010
2. Cumulative flow for F0101 = Flow for F0101+ Flow for F0010
3. Cumulative flow for F0012 = Flow for F0012
4. Cumulative flow for F0014 = Flow for F0014 + Flow for F0012 + Cumulative flow for F0101
5. Cumulative flow for F0016 = Flow for F0016 + Cumulative flow for F0014
6. Cumulative flow for F0022 = Flow for F0022 + Cumulative flow for F0016

The most downstream site at F0022 carries 100% of the river flow (cumulative). The proportion of flow recorded as compared to F0022 is calculated for all other sites.

4.2.2 Dry and wet weather flow:

We segregate the flow data into dry and wet weather flows based on daily rainfall. Four nearest radar points are found at each flow site, and the average rainfall at those points is considered the rainfall for that flow site. Dry days are defined as the days for which the daily rainfall

depth $<$1mm, and wet days are days when rainfall depth $\geq$ 1mm. The threshold of 1mm for classifying dry days has been frequently used in many studies including Rivoire et al. [2019] and Zolina et al. [2010]. Of the 381 days, 225 were classified as dry, while 156 were classified as wet.

After classifying the flow for dry and wet weather, we again compute and compare the cumulative flow for both weather conditions.

4.2.3 Land cover data:

Land cover indicates the physical land type of the Earth's surface for a given region. This includes vegetation, human construction, water, and bare ground. We obtain Bradford's land cover raster data from ESRI's living atlas of the world [ESRI, n.d.] collection. After obtaining the data, we use GIS to calculate the area and proportion of different land cover types in each flow gauge catchment. We will compare this to the changes in flow accumulation during dry and wet periods to draw insights about the runoff rates for all site catchments. ESRI specifies seven different land cover types:

- **Water:** Areas where water was predominantly present throughout the year.
- **Trees:** Any significant clustering of tall (15m or higher) dense trees.
- **Flooded Vegetation:** Any vegetation areas with obvious intermixing of water throughout the year.
- **Crops:** Human planted cereals, grasses and crops not at tree height.
- **Built Area:** Human-made structures, including impervious surfaces like parking lots, housing, and buildings.
- **Bare ground:** Areas of rock or soil with little to no vegetation throughout the year.
- **Rangelands:** Open areas covered in homogeneous grasses with short vegetation.

4.3 Results

4.3.1 Flow accumulation

The average flow accumulation and the percentage contribution to flow for each site are summarised below.

***Table 4.1*:** Table showing each site's average cumulative flow values and their proportion to the total flow.

Site	Avg Cumulative flow (L/s)	Percentage contribution
F0012	16.91	0.55%
F0010	590.63	19.14%
F0101	1209.08	39.19%
F0014	1943.81	63%
F0016	2465.88	79.92%
F0022	3085.28	100%

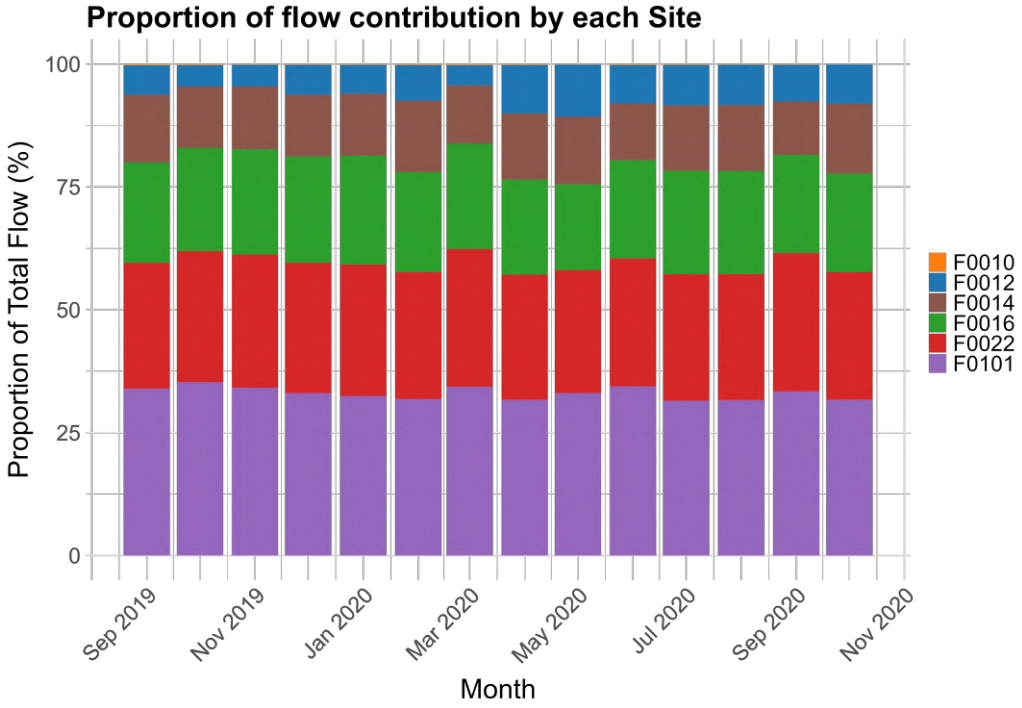


***Figure 4.3*:** Stacked bar plot showing the proportion of each flow site in total flow per month. The length of each bar segment represents the proportion of flow contributed by that site.

4.3.2 Dry and wet weather flows

A summary of flow accumulation analysis for dry and wet weather conditions is shown below.

***Table 4.2*:** Summary of flow accumulation analysis during dry weather conditions

Site	Avg Cumulative Flow (L/s)	Percentage contribution
F0012	10.79	0.49%
F0010	433.45	19.71%
F0101	829.11	37.7%
F0014	1423.67	64.73%
F0016	1799.16	81.8%
F0022	2199.45	100%

***Table 4.3*:** Summary of flow accumulation analysis during wet weather conditions

Site	Avg Cumulative Flow (L/s)	Percentage contribution
F0012	25.75	0.59%
F0010	817.35	18.73%
F0101	1757.14	40.27%
F0014	2694.05	61.75%
F0016	3427.54	78.56%
F0022	4362.98	100%

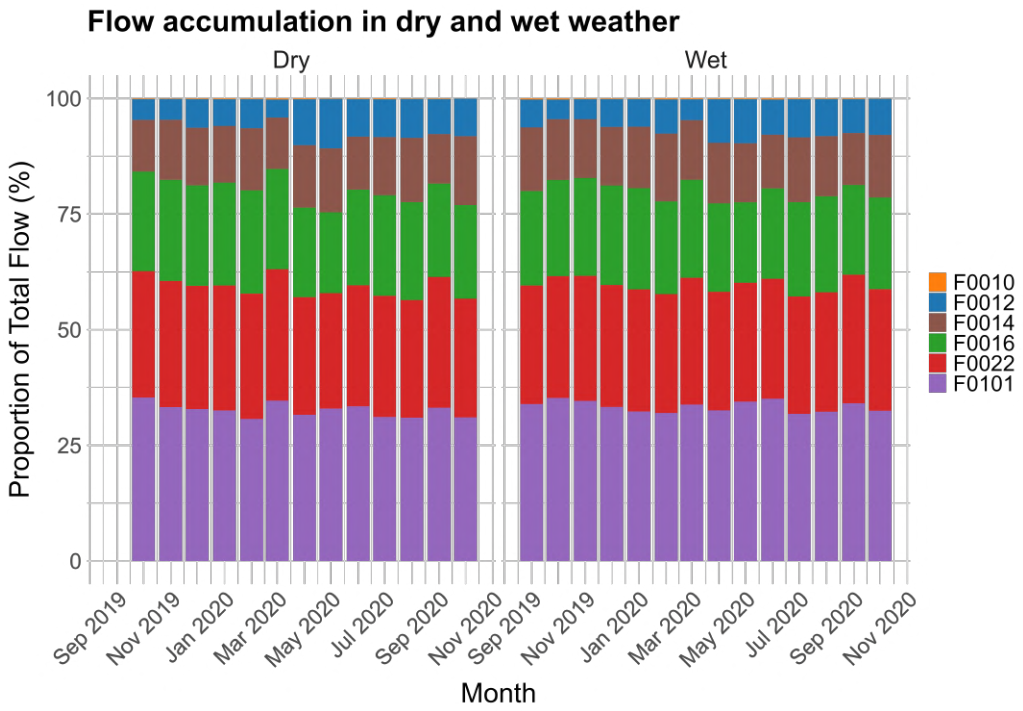


***Figure 4.4*:** Stacked bar plot showing the proportion of each flow site in total flow per month for dry and wet weathers. The length of each bar segment represents the proportion of flow contributed by that site.

4.3.3 Land cover analysis

Table 4.4 shows the average increase in flow accumulation at each site in wet weather compared to dry weather per unit catchment area. Flow, in this case, is the regular discharge recorded at each site and not the cumulative flow. It can be seen that flow per unit catchment area increases the fastest for sites F0014 and F0022 during rainfall, implying a high surface runoff rate likely due to highly urbanised catchment areas. The runoff rate is not uniform across all flow sites.

***Table 4.4*:** Average increase in wet weather flow for all sites per unit catchment area. Flow is the actual flow recorded and not the cumulative flow.

Site	**Avg dry flow (L/s)**	**Avg wet flow (L/s)**	**Δflow**	**Catchment area (km^2)**	**Δflow/Catchment area**
F0012	10.786	25.746	14.959	4.54	3.29
F0010	433.446	817.350	383.904	19.67	19.51
F0101	395.6626	939.794	544.131	19.67	27.66
F0014	583.773	911.155	327.382	6.15	53.23
F0016	375.487	733.492	358.005	12.46	28.73
F0022	400.295	935.441	535.146	9.5	56.33

Figure 4.5 shows the land cover map for the whole river catchment.

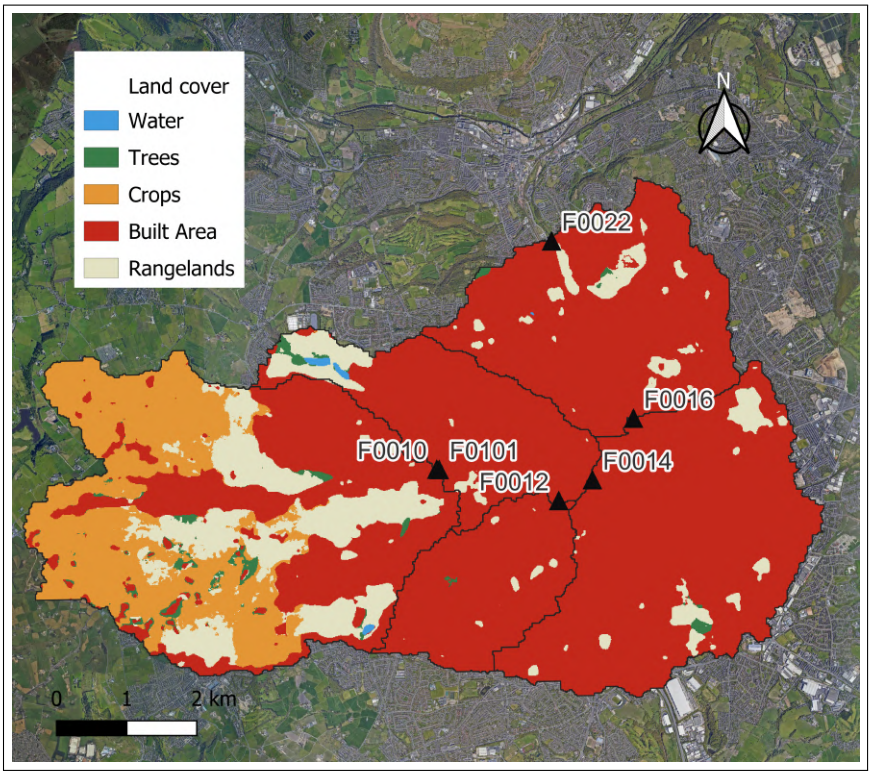


***Figure 4.5*:** Land cover map for all flow gauge catchments.

The following table summarises the area and land cover percentages for all flow gauge catchments.

***Table 4.5*:** Table showing the land cover types in each catchment area and their respective areas and percentages.

Site	**Land cover**	**Area (sq. km)**	**Percentage**
F0010 and F0101	Water	0.5719	0.357%
	Trees	8.6762	5.416%
	Flooded Vegetation	0.0275	0.017%
	Crops	41.2787	25.769%
	Built Area	86.9317	54.268%
	Bare ground	0.0067	0.004%
	Rangelands	22.6961	14.168%
F0012	Trees	0.0106	0.234%
	Built Area	4.4564	98.211%
	Rangelands	0.0706	1.556%
F0014	Water	0.058	0.943%
	Trees	0.0822	1.337%
	Crops	0.0013	0.021%
	Built Area	5.294	86.098%
	Rangelands	0.7133	11.601%
F0016	Trees	0.0493	0.396%
	Built Area	11.765	94.487%
	Rangelands	0.6372	5.117%
F0022	Water	0.0017	0.018%
	Trees	0.0378	0.398%
	Built Area	8.7237	91.93%
	Bare ground	0.0001	0.001%
	Rangelands	0.7262	7.653%

Chapter 5

Transit time calculation

Computing the transit time of pollution for various reaches of the river helps in identifying the source of the pollution by tracking the pollution peak downstream. It also gives insights into sediment transport's dynamics and seasonal variability across the channel. The RiverSpill model developed by Samuels et al. [2006] simulates the transport of contaminants in river systems by calculating their travel time and evaluating how long it takes for a spill to reach and potentially contaminate drinking water supply stations.

For the purposes of this section, we will only consider total ammonia nitrogen (TAN) as the primary river pollutant and identifier of water quality status.

5.1 Methodology

5.1.1 Indentifying water quality events

The first step is to find the significant water quality events for the time period. To do this, we use the 'moderate' water quality threshold[1] for total ammonia as defined by the Water Framework directive [European Commission, 2024], which is 0.75 mg/L. The heatmap in figure 5.1 shows the number of quality monitoring sites for which the threshold of 0.75 mg/L was exceeded on the same day.

[1] 'Good' threshold of 0.3 mg/L was being violated nearly 100% of the times.

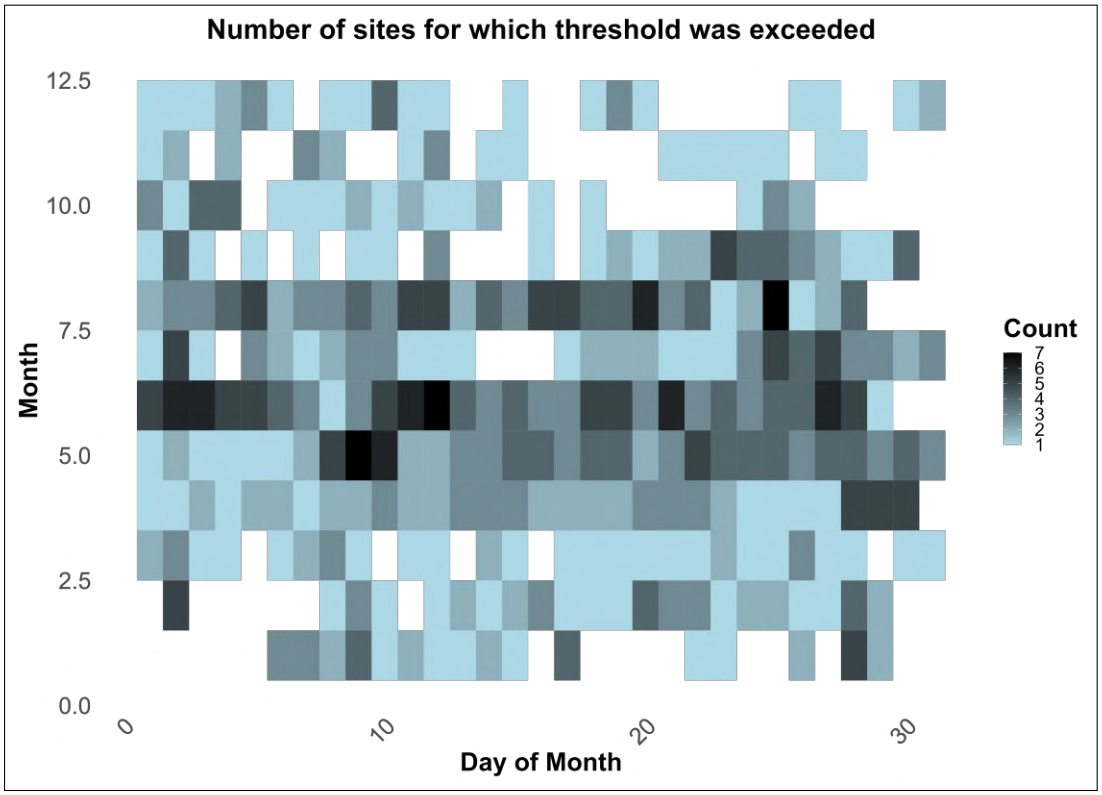


***Figure 5.1*:** Heatmap showing the number of quality sites for which the threshold of 0.75 mg/L was exceeded on the same day.

For calculating transit times, we will consider the quality monitors located in the Bradford beck only; S0014, S0015, S0016, S0022, and S0024. There are 24 dates on which the total ammonia threshold was exceeded at all five monitoring sites. We will only be showing the calculation for four of those dates: **29th April 2020**, **9th May 2020**, **3rd June 2020**, and **2nd July 2020**.

5.1.2 Transit time and speed calculation

Transit times are determined by calculating the time difference between peak TAN measurements for each quality site pair. Identifying peak TAN measurements is cumbersome since multiple minor events can co-occur with different measured TAN peaks. We need to ensure that we calculate the time difference between peaks that correspond to the same quality event. Therefore, we first plot the TAN values and find the distinct peaks in the plot using the `find_peaks` functionality of the `scipy` library in python [2]. After assuring that the peak TAN value in all quality monitors coincides with the same quality event, we find the time difference to calculate transit time.

To find the speed of pollution, we need to measure the distance between each pair of monitoring sites. The distance between each site is computed using QGIS, which provides built-in

[2] `find_peaks` function returns all the local maxima of a 1D array by comparing neighbourhood values.

support for finding the length of river geometries.

***Table 5.1*:** Distance between each pair of quality monitoring sites.

Quality site	**Length (in m)**
S0014 - S0015	93.45
S0015 - S0016	1550.04
S0016 - S0022	2975.72
S0022 - S0024	1563.08

Pollution speed is computed by dividing the distance between each site by the respective transit time.

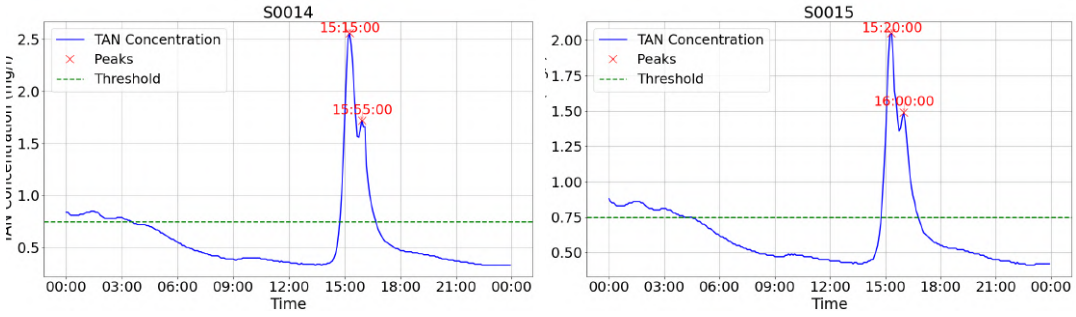
$$\text{Speed} = \frac{\text{Distance}}{\text{Transit Time}}$$

A complete pipeline for peak identification and transit time calculation for quality events is made available in Python.

5.2 Results

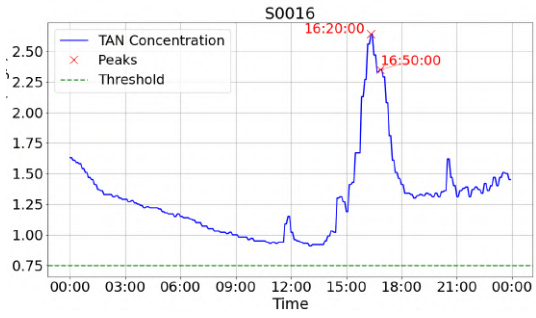
5.2.1 Event of 29th April 2020

The plot of TAN concentration (with peaks) on the event date for all five sites is shown below.

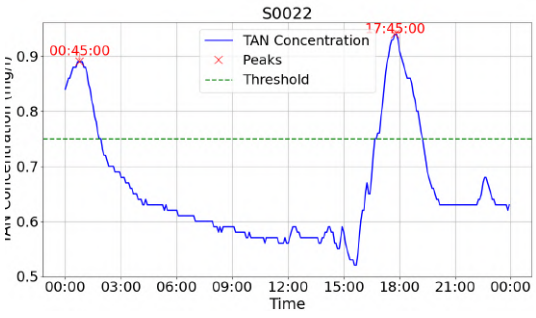


(a) Site S0014 (Peak TAN = 2.56mg/L)

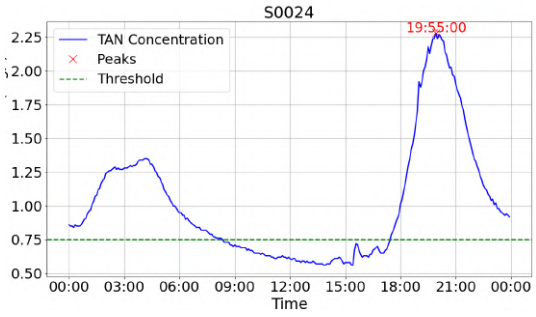
(b) Site S0015 (Peak TAN = 2.05mg/L)



(c) Site S0016 (Peak TAN = 2.64mg/L)



(d) Site S0022 (Peak TAN = 0.94mg/L)



(e) Site S0024 (Peak TAN = 2.28mg/L)

***Figure 5.2*:** TAN concentration with peak times for all quality sites on 29-04-2020.

The calculated transit times and velocities are shown in the following table.

***Table 5.2*:** Transit times and velocities of pollution on the event day.

Site ID	**Transit time (mins)**	**Velocity (m/s)**
S0014 - S0015	5.0	0.31
S0015 - S0016	60.0	0.43
S0016 - S0022	85.0	0.58
S0022 - S0024	130.0	0.2

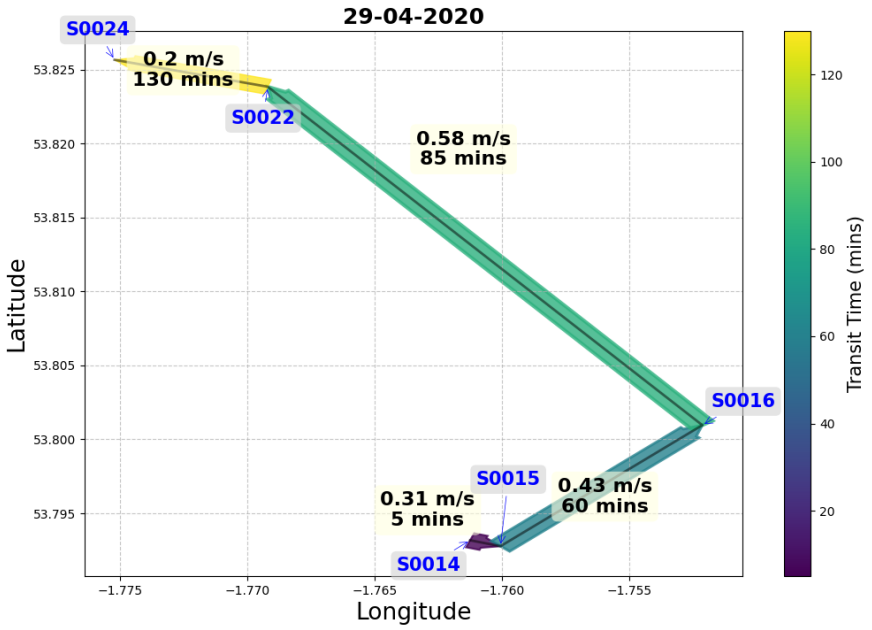


***Figure 5.3*:** Plot showing the transit times and velocities at each site in the river.

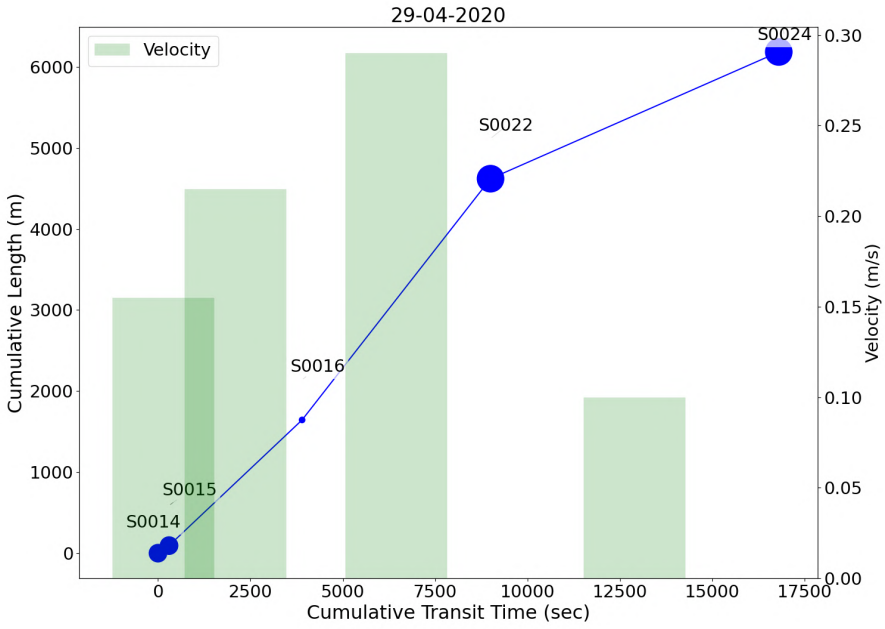
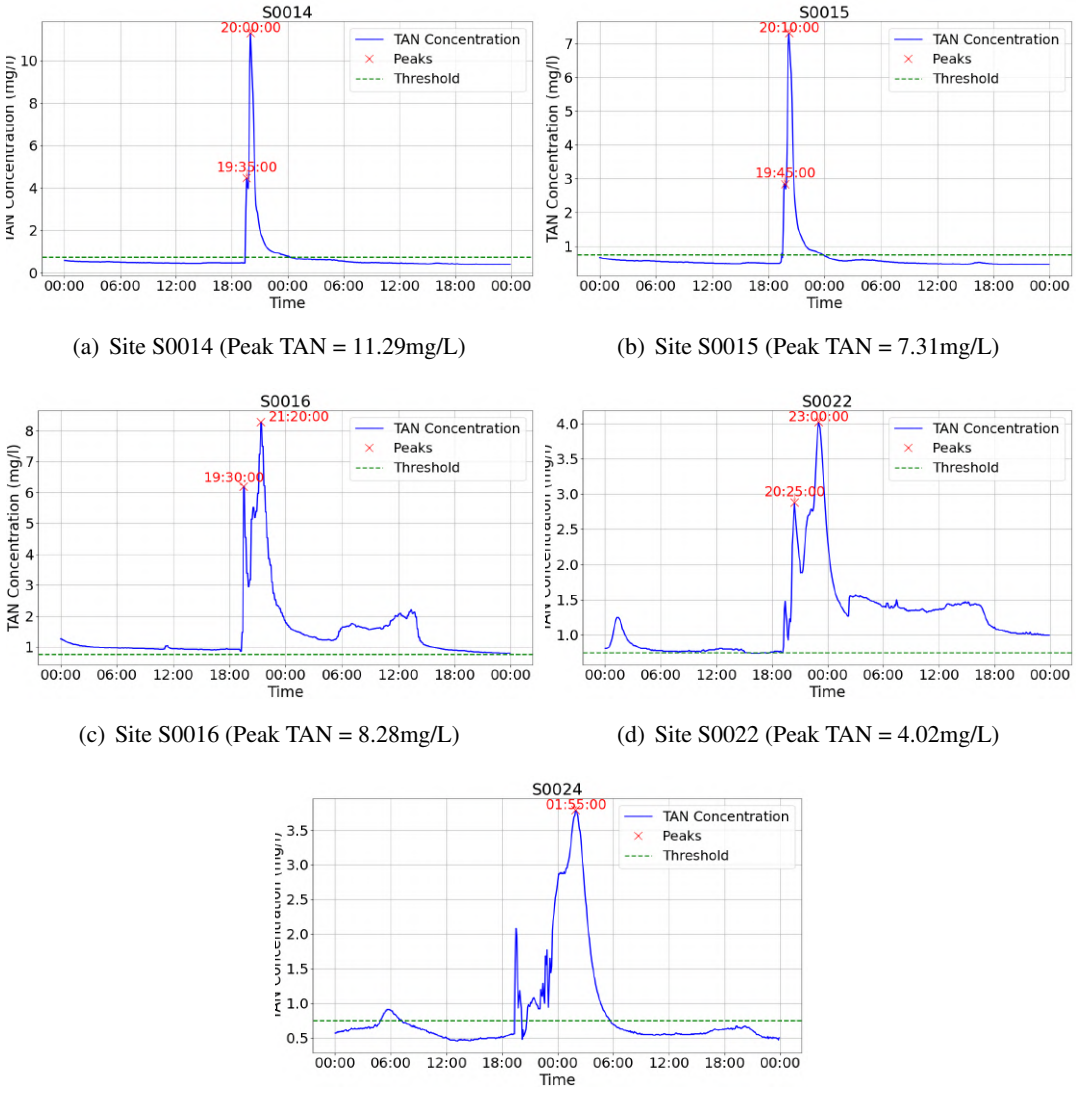


***Figure 5.4*:** Plot showing the cumulative length vs cumulative transit time for the event of 29th April. The slope of each line segment gives the speed of pollution between the two sites and the size of each point is proportional to the flow measured at that site.

5.2.2 Event of 9th May 2020

The plot of TAN concentration (with peaks) on the event date for all five sites is shown below.



(a) Site S0014 (Peak TAN = 11.29mg/L)

(b) Site S0015 (Peak TAN = 7.31mg/L)

(c) Site S0016 (Peak TAN = 8.28mg/L)

(d) Site S0022 (Peak TAN = 4.02mg/L)

(e) Site S0024 (Peak TAN = 3.79mg/L)

***Figure 5.5*:** TAN concentration with peak times for all quality sites on 09-05-2020.

The calculated transit times and velocities are shown in the following table.

***Table 5.3*:** Transit times and velocities of pollution on the event day.

Site ID	Transit time (mins)	Velocity (m/s)
S0014 - S0015	10.0	0.16
S0015 - S0016	70.0	0.37
S0016 - S0022	100.0	0.50
S0022 - S0024	175.0	0.15

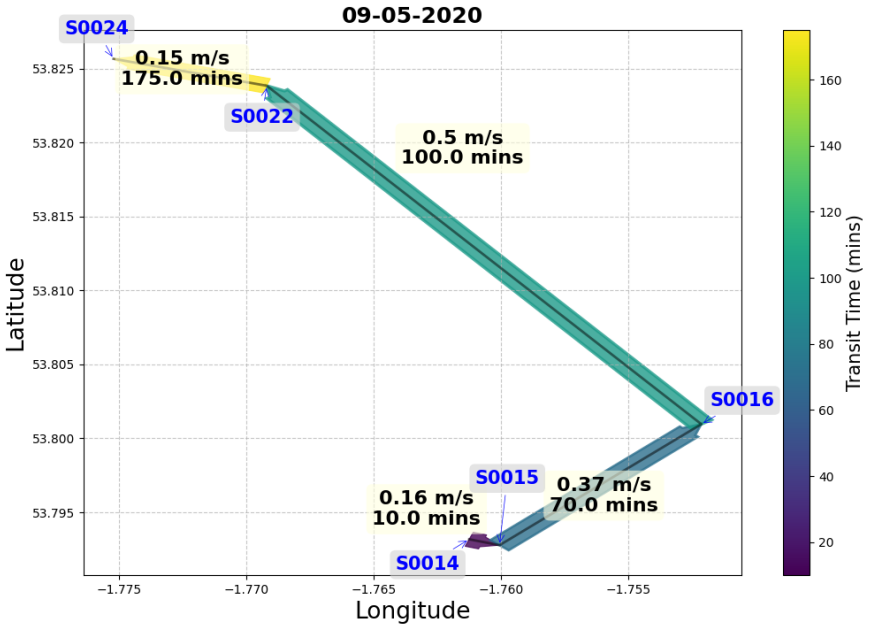


***Figure 5.6*:** Plot showing the transit times and velocities at each site in the river.

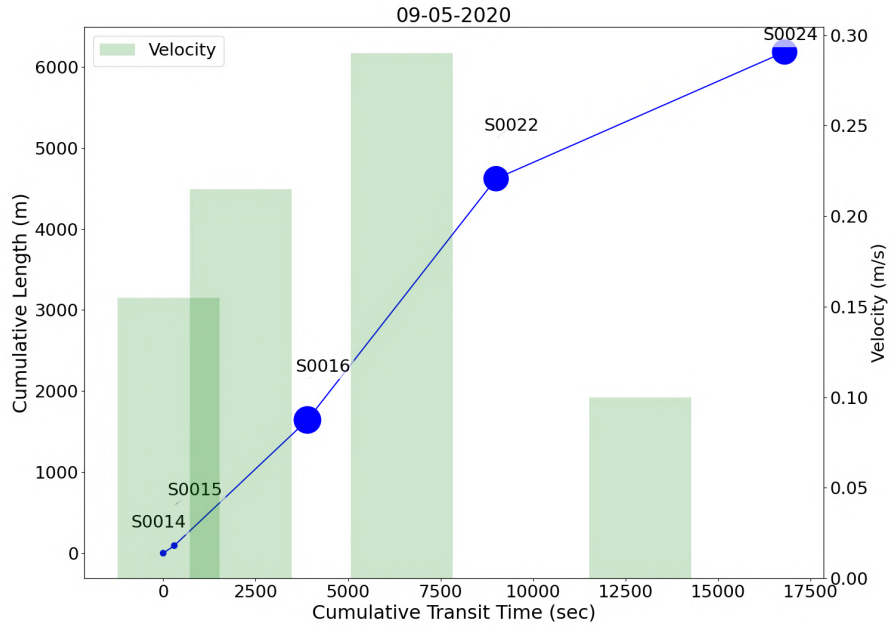
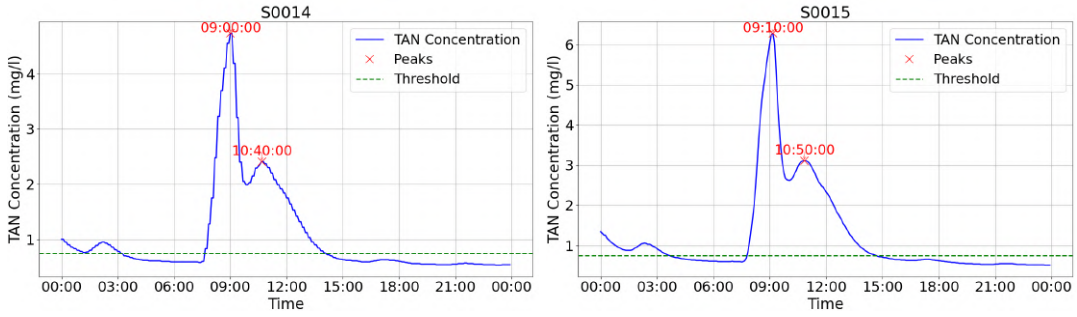


***Figure 5.7*:** Plot showing the cumulative length vs cumulative transit time for the event of 9th May. The slope of each line segment gives the speed of pollution between the two sites and the size of each point is proportional to the flow measured at that site.

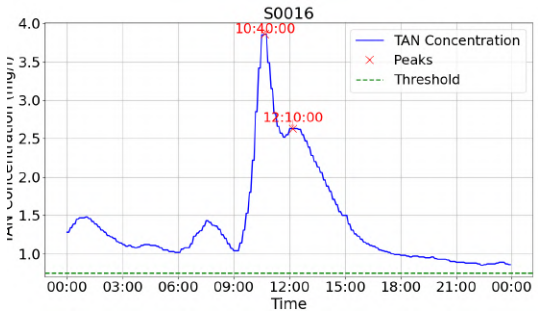
5.2.3 Event of 3rd June 2020

The plot of TAN concentration (with peaks) on the event date for all five sites is shown below.

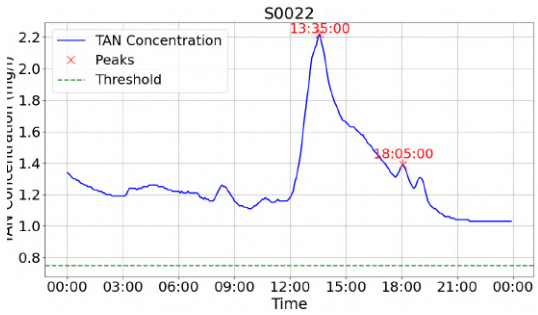


(a) Site S0014 (Peak TAN = 2.41mg/L)

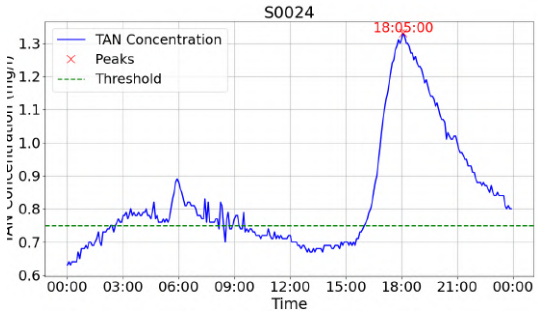
(b) Site S0015 (Peak TAN = 3.11mg/L)



(c) Site S0016 (Peak TAN = 2.63mg/L)



(d) Site S0022 (Peak TAN = 2.22mg/L)



(e) Site S0024 (Peak TAN = 1.33mg/L)

***Figure 5.8*:** TAN concentration with peak times for all quality sites on 03-06-2020. For transit time calculation, we take the second-highest peak for sites S0014, S0015, and S0016.

The calculated transit times and velocities are shown in the following table.

***Table 5.4*:** Transit times and velocities of pollution on the event day.

Site ID	**Transit time (mins)**	**Velocity (m/s)**
S0014 - S0015	10.0	0.16
S0015 - S0016	80.0	0.32
S0016 - S0022	85.0	0.58
S0022 - S0024	270.0	0.1

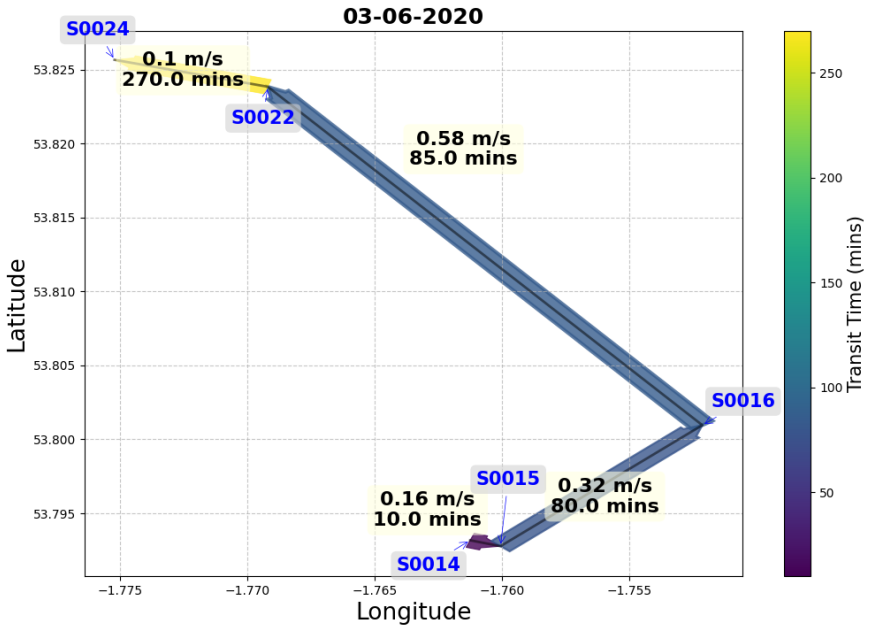


***Figure 5.9*:** Plot showing the transit times and velocities at each site in the river.

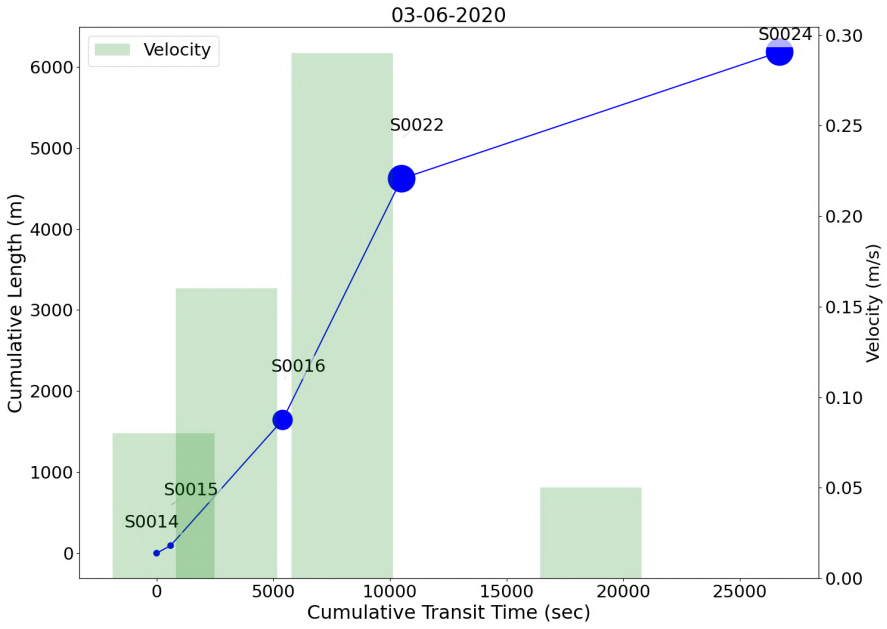
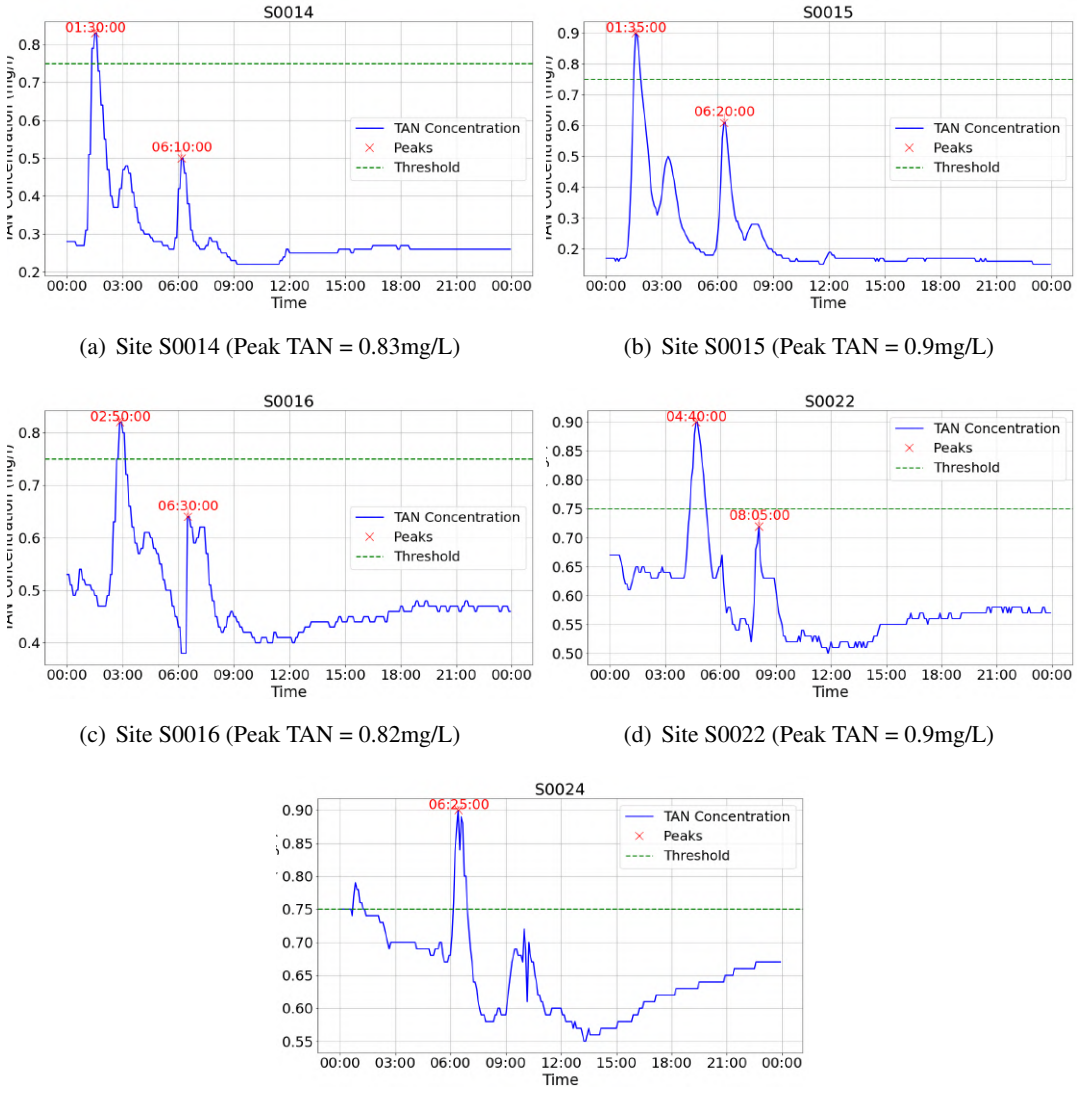


***Figure 5.10*:** Plot showing the cumulative length vs cumulative transit time for the event of 3rd June. The slope of each line segment gives the speed of pollution between the two sites and the size of each point is proportional to the flow measured at that site.

5.2.4 Event of 2nd July 2020

The plot of TAN concentration (with peaks) on the event date for all five sites is shown below.



(a) Site S0014 (Peak TAN = 0.83mg/L)

(b) Site S0015 (Peak TAN = 0.9mg/L)

(c) Site S0016 (Peak TAN = 0.82mg/L)

(d) Site S0022 (Peak TAN = 0.9mg/L)

(e) Site S0024 (Peak TAN = 0.9mg/L)

***Figure 5.11*:** TAN concentration with peak times for all quality sites on 02-07-2020.

The calculated transit times and velocities are shown in the following table.

***Table 5.5*:** Transit times and velocities of pollution on the event day.

Site ID	Transit time (mins)	Velocity (m/s)
S0014 - S0015	5.0	0.31
S0015 - S0016	75.0	0.34
S0016 - S0022	110.0	0.45
S0022 - S0024	105.0	0.25

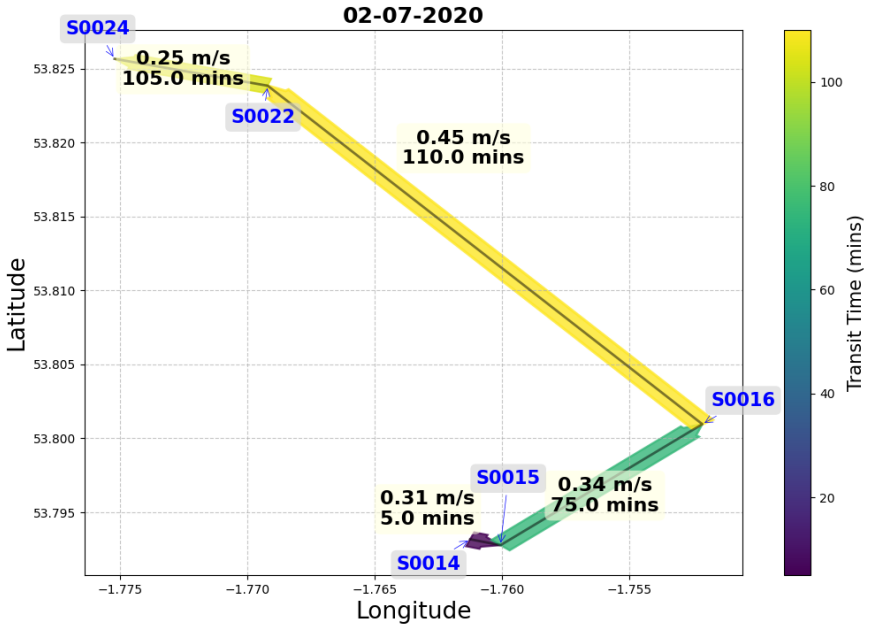


***Figure 5.12*:** Plot showing the transit times and velocities at each site in the river.

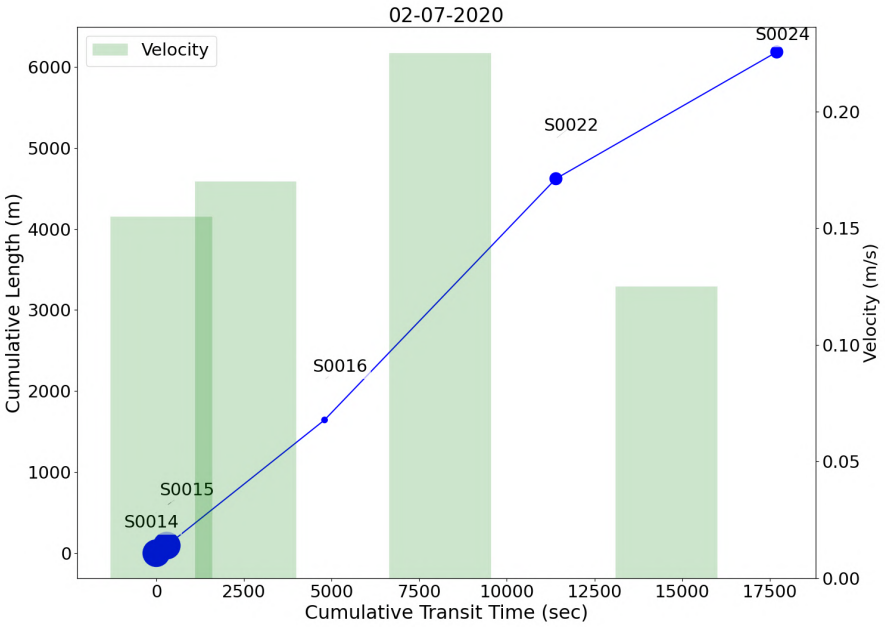


***Figure 5.13*:** Plot showing the cumulative length vs cumulative transit time for the event of 2nd July. The slope of each line segment gives the speed of pollution between the two sites and the size of each point is proportional to the flow measured at that site.

Chapter 6

Conclusions and discussion

6.1 Conclusions

6.1.1 EDA plots

- The flow plots reveal that in most parts of the catchment, the flow is extremely 'flashy', i.e. changes rapidly with rainfall.
- The Water quality status of the river is inferior, and TAN follows an upward trend for the time period.
- Acidity in the river negatively correlates with flow volume.
- The Catchment area for site S0010 is less urbanised as it displays low surface runoff rates.

6.1.2 Flow accumulation analysis

- Interestingly, flow for the most downstream site, F0022, is not the highest. Rather, flow at F0014 is the highest, possibly due to the merging of many tributaries at F0014.
- Flow accumulation change per catchment area is an important metric which reveals that the flow accumulation does not change due to uniform runoff rates. Urbanised parts of the catchment have a higher runoff rate. The accumulation for the flow site F0016, located in the city centre, is less, likely due to the diversion tunnel.
- Land cover analysis and flow accumulation can be further analysed to draw insights into runoff rates and catchment characteristics.

6.1.3 Transit times

- May, June and August witness the highest pollution events as seen from figure 5.1.
- Pollution speeds are variable and depend on the flow of the river.

- The peak value of TAN at each site can be used to trace the source of the pollution event.

6.2 Discussion

- Flow recorded at the most downstream site is not the highest, likely because of diversion and groundwater seepage along the river.
- DEM data used for catchment delineation has a resolution of 30m, which might affect the results. For instance, the catchment areas of F0010 and F0101 are calculated to be the same because the sites are very close together, and the resolution might not be high enough to capture the differences in locations.
- Cumulative flow is used instead of raw flow values for flow accumulation. This is done to ensure consistency in the results since the flows are highly variable, and flow accumulation using highly variable flows produces erroneous results. Cumulative flow also validates the theoretical result that flow at each downstream flow site is a sum of the flows from the upstream sites.
- Pollution speeds calculated using peak time difference are different from river flow velocities. This is because as ammonia (TAN) travels downstream, it mixes and disperses with the surrounding water due to turbulence, which results in slower travel times.
- There is no way to ascertain that different TAN peaks correspond to the same pollution event.
- There were no common pollution events between sites S0101 and S0014. Hence, the transit times and speeds could not be computed for that stretch of the river.

Bibliography

Dhanya Pramod Anil Jadhav and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019. doi: 10.1080/08839514.2019.1637138. URL https://doi.org/10.1080/08839514.2019.1637138.

Jeffrey G. Arnold, Raghavan Srinivasan, Raghavan S. Muttiah, and Jimmy R. Williams. Large area hydrologic modeling and assessment part i: Model development. *Journal of the American Water Resources Association*, 34(1):73–89, 1998. doi: 10.1111/j.1752-1688.1998.tb05961.x. URL https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

BBC News, E. S. J. F. A. S. Woodcock, and B. D. A. L. Rogers. BBC uncovers 6,000 possible illegal sewage spills in one year. *BBC News*, June 13 2024. URL https://www.bbc.co.uk/news/articles/c4nn46rjej6o.

Department for Environment, Food & Rural Affairs et al. New plan for cleaner and more plentiful water. https://www.gov.uk/government/news/new-plan-for-cleaner-and-more-plentiful-water, 2023. Accessed: 30 August 2024.

DHI. *MIKE 11 - A Modeling System for Rivers and Channels, Reference Manual*. DHI Water & Environment, Hørsholm, Denmark, 2007.

ESRI. Arcgis living atlas of the world, n.d. URL https://livingatlas.arcgis.com/en/home/. Accessed: 2024-09-02.

European Commission. Water framework directive. https://environment.ec.europa.eu/topics/water/water-framework-directive, 2024. Accessed: 2024-08-31.

Ramu Gautam and Shahram Latifi. Comparison of simple missing data imputation techniques for numerical and categorical datasets. *Journal of Research in Engineering and Applied Sciences*, 8:468–475, 04 2023. doi: 10.46565/jreas.202381468-475.

Tracey H. Goodwin, Andrew R. Young, Matthew G.R. Holmes, Gareth H. Old, Ned Hewitt, Graham J.L. Leeks, John C. Packman, and Barnaby P.G. Smith. The temporal and spatial vari-

ability of sediment transport and yields within the bradford beck catchment, west yorkshire. *Science of The Total Environment*, 314-316:475–494, 2003. ISSN 0048-9697. doi: https://doi.org/10.1016/S0048-9697(03)00069-X. URL https://www.sciencedirect.com/science/article/pii/S004896970300069X. Land Ocean Interaction: processes, functioning and environmental management:a UK perspective.

Fatemeh Hashemi, Ina Pohle, Johannes W. M. Pullens, Henrik Tornbjerg, Katarina Kyllmar, Hannu Marttila, Ahti Lepistö, Bjørn Kløve, Martyn Futter, and Brian Kronvang. Conceptual mini-catchment typologies for testing dominant controls of nutrient dynamics in three nordic countries. *Water*, 12:1776, 2020. doi: 10.3390/w12061776. URL https://doi.org/10.3390/w12061776.

Iqbal Hossain and Monzur Imteaz. Catstream: An integrated catchment-stream water quality model. 01 2013.

Derek Karssenberg, Oliver Schmitz, Peter Salamon, Kor de Jong, and Marc F.P. Bierkens. A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling & Software*, 25(4):489–502, 2010. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2009.10.004. URL https://www.sciencedirect.com/science/article/pii/S1364815209002643.

Met Office, D. Hollis, M. McCarthy, M. Kendon, T. Legg, and I. Simpson. HadUK-Grid Gridded Climate Observations on a 1km grid over the UK, v1.0.3.0 (1862-2020). https://dx.doi.org/10.5285/786b3ce6be54468496a3e11ce2f2669c, 2021. Accessed: 08 September 2021.

NASA Earth Science Data Systems (ESDS). SRTM — EarthData, March 4 2024. URL https://www.earthdata.nasa.gov/sensors/srtm. Accessed: 2024-09-01.

Jock O'Callaghan and David M. Mark. The extraction of drainage networks from digital elevation data. *Comput. Vis. Graph. Image Process.*, 28:323–344, 1984. URL https://api.semanticscholar.org/CorpusID:32850139.

Gareth H. Old, Graham J.L. Leeks, John C. Packman, Barnaby P.G. Smith, Scott Lewis, and Edward J. Hewitt. River flow and associated transport of sediments and solutes through a highly urbanised catchment, bradford, west yorkshire. *Science of The Total Environment*, 360(1):98–108, 2006. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2005.08.028. URL https://www.sciencedirect.com/science/article/pii/S0048969705005632. Urban Environmental Research in the UK: The Urban Regeneration and the Environment (NERC URGENT) Programme and associated studies.

PCRaster. *lddcreate — PCRaster 4.4.1 Documentation*, n.d. URL https://pcraster.geo.uu.nl/pcraster/4.4.1/documentation/pcraster_manual/sphinx/op_lddcreate.html. Accessed: 2024-09-01.

P. Rivoire, Y. Tramblay, L. Neppel, E. Hertig, and S. M. Vicente-Serrano. Impact of the dry-day definition on mediterranean extreme dry-spell analysis. *Natural Hazards and Earth System Sciences*, 19(8):1629–1638, 2019. doi: 10.5194/nhess-19-1629-2019. URL https://nhess.copernicus.org/articles/19/1629/2019/.

William Samuels, David Amstutz, Rakesh Bahadur, and Jonathan Pickus. Riverspill : A national application for drinking water protection. *Journal of Hydraulic Engineering-asce - J HYDRAUL ENG-ASCE*, 132, 04 2006. doi: 10.1061/(ASCE)0733-9429(2006)132:4(393).

A. Strahler. Quantitative geomorphology of drainage basins and channel networks. In V. Chow, editor, *Handbook of Applied Hydrology*, pages 439–476. McGraw-Hill, New York, 1964.

Viet-Bach Tran, Hiroshi Ishidaira, Takashi Nakamura, Thu-Nga Do, and Kei Nishida. Estimation of nitrogen load with multi-pollution sources using the swat model: a case study in the cau river basin in northern vietnam. *Journal of Water and Environment Technology*, 15(3): 106–119, 2017. doi: 10.2965/jwet.16-052.

UNESCO World Water Assessment Programme. *The United Nations World Water Development Report, 2017: Wastewater: The Untapped Resource*. UNESCO, Paris, 2017. ISBN 978-92-3-100201-4. URL https://www.unesco.org/en/dynamic-content/wwdr-2017. 180 p., illus., maps.

U.S. Environmental Protection Agency. Implementation guidance on surface water quality standards for nutrients, 2012. URL https://www.chiwater.com/Files/Swmm_Apps_Manual.pdf. Accessed: 2024-08-01.

T. Van Emmerik, S. De Lange, R. Frings, L. Schreyers, H. Aalderink, J. Leusink, F. Begemann, E. Hamers, R. Hauk, N. Janssens, P. Jansson, N. Joosse, D. Kelder, T. Van Der Kuijl, R. Lotcheris, A. Löhr, Y. Mellink, R. Pinto, P. Tasseron, and P. Vriend. Hydrology as a driver of floating river plastic transport. *Earth's Future*, 10(8), 2022. doi: 10.1029/2022ef002811. URL https://doi.org/10.1029/2022ef002811.

Lei Wang, Deyang Dang, Yusheng Liu, Xiaobo Peng, and Ruisheng Liu. Dynamic water environment capacity assessment based on control unit coupled with SWAT model and differential evolution algorithm. *Water*, 15(10):1817, 2023. doi: 10.3390/w15101817. URL https://doi.org/10.3390/w15101817.

WHO & UNICEF JMP. *Progress on Household Drinking Water, Sanitation and Hygiene 2000-2020: Five Years into the SDGs*. WHO and UNICEF, Geneva, 2021. Report.

Olga Zolina, Clemens Simmer, Sergey K Gulev, and Stefan Kollet. Changing structure of european precipitation: Longer wet periods leading to more abundant rainfalls. *Geophysical Research Letters*, 37(6), 2010. doi: 10.1029/2010gl042468. URL https://doi.org/10.1029/2010gl042468.

Appendix A

Flow and quality monitoring locations

The following table summarises the locations of the flow and quality monitoring sites.

***Table A.1*:** Details of all flow and quality monitoring sites

Site code	IETG code	Monitoring type	Watercourse	Location	Latitude	Longitude
BB10a	S0010	Water quality	Middlebrook	Upstream Chellow Dene Beck	53.794872	-1.794887
BB10a	F0010	Water flow	Middlebrook	Upstream Chellow Dene Beck	53.794872	-1.7948875
BB10b	F0101	Water flow	Bradford Beck	Downstream Chellow Dene Beck	53.794827	-1.794374
BB10b	S0101	Water quality	Bradford Beck	Downstream Chellow Dene Beck	53.79482931	-1.79431783
BB12	F0012	Water flow	Westbrook	Theatre in the Mill, University of Bradford	53.7905803	-1.768621
BB14	S0014	Water quality	Bradford Beck	Upstream Westbrook	53.793175	-1.761259
BB14	F0014	Water flow	Bradford Beck	Upstream Westbrook	53.793175	-1.761259
BB15	S0015	Water quality	Bradford Beck	Downstream Westbrook	53.792774	-1.760066
BB16	S0016	Water quality	Bradford Beck	Amber Mill car park Bradford	53.800961	-1.752136
BB16	F0016	Water flow	Bradford Beck	Amber Mill car park Bradford	53.800961	-1.752136
BB22	S0022	Water quality	Bradford Beck	VPG Group car park Airedale House	53.823853	-1.769213
BB22	F0022	Water flow	Bradford Beck	VPG Group car park Airedale House	53.823853	-1.769213
BB27	S0027	Water quality	Red Beck	Ground of Acorn Stairlifts Norwood Avenue	53.825672	-1.775231
BB24	S0024	Water quality	Bradford Beck	Near AWP Construction Leeds Road Shipley	53.835377	-1.770327

Appendix B

KNN imputer

B.1 Working of KNN Imputer:

Given a dataset $X \in \mathbb{R}^{m \times n}$, where m is the number of samples and n is the number of features, suppose we have a missing value x_{ij} in the dataset, where i is the sample index and j is the feature index.

1. **Standard Scaling**

   KNN depends on the distance between data points. So before performing imputation, we apply standard scaling [1] to the dataset. Each feature j in the dataset is scaled to have a mean of 0 and a standard deviation of 1. The scaling transformation for a feature j is given by:

   $$x_{ij}^{\text{scaled}} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

   where μ_j and σ_j are the mean and standard deviation of the feature j across all samples that have non-missing values.

2. **Distance Calculation**

   For each sample X_i (i-th row) with a missing value at feature j, we calculate the distance[2] to all other samples X_k (k-th row) that have non-missing values for feature j. The Euclidean distance between X_i and X_k is calculated using the scaled data as:

   $$d(X_i^{\text{scaled}}, X_k^{\text{scaled}}) = \sqrt{\sum_{l=1, l \neq j}^{n} (x_{il}^{\text{scaled}} - x_{kl}^{\text{scaled}})^2}$$

   This sum excludes the feature j because x_{ij} is missing.

[1] There are other scaling techniques like min-max scaling and normalization, which can also be used.

[2] various distance metrics can be used to define the "closeness" of points. Some examples are Manhattan distance, Minkowski distance, Chebyshev distance, etc.

3. **Finding Nearest Neighbors**

   After computing the distances on the scaled data, we select the k smallest distances. Let's denote the indices of these k nearest neighbors as $\{k_1, k_2, \ldots, k_k\}$.

4. **Imputing Missing Value**

   The missing value x_{ij} is imputed using the mean value[3] of the nearest neighbours:

$$x_{ij}^{\text{scaled}} = \frac{1}{k} \sum_{p=1}^{k} x_{k_p j}^{\text{scaled}}$$

   Finally, the imputed value is transformed back to the original scale by reversing the standard scaling transformation:

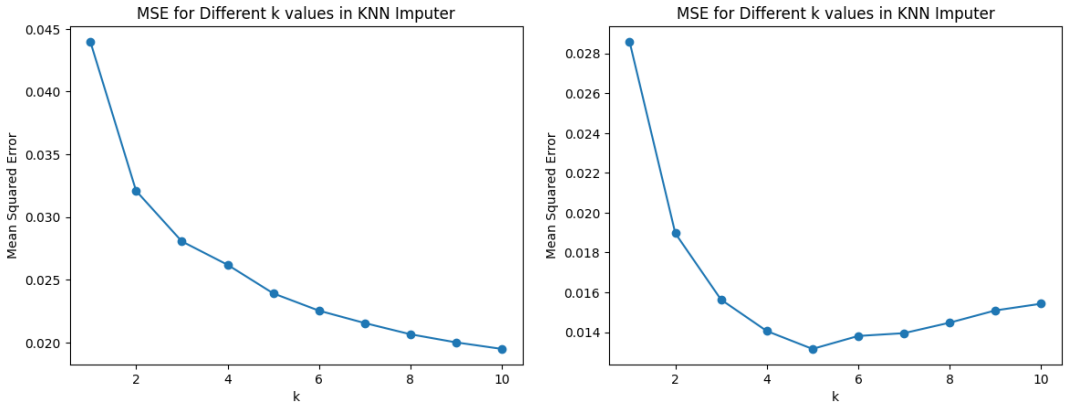
$$x_{ij} = x_{ij}^{\text{scaled}} \times \sigma_j + \mu_j$$

   where μ_j and σ_j are the mean and standard deviation used during the scaling process.

B.2 Plots

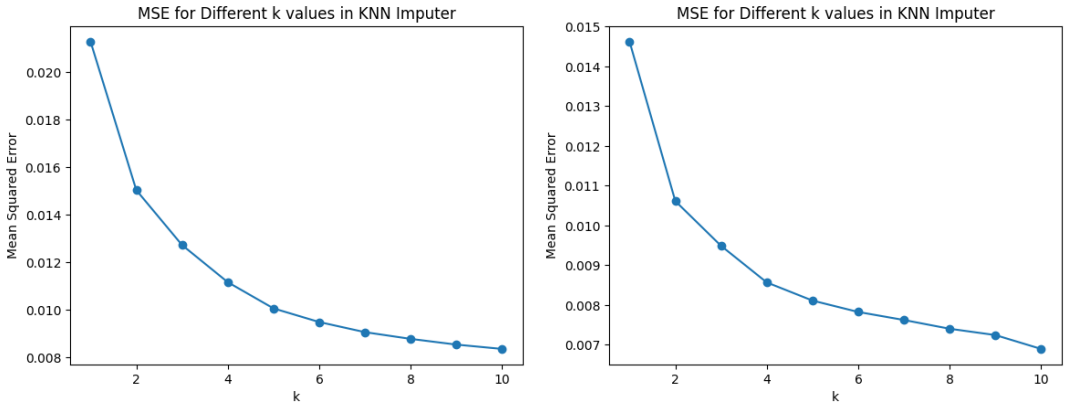
The plots of MSE vs k for all datasets are shown below.

[3] Other aggregation techniques like median, mode, and weighted mean can also be used instead of mean.



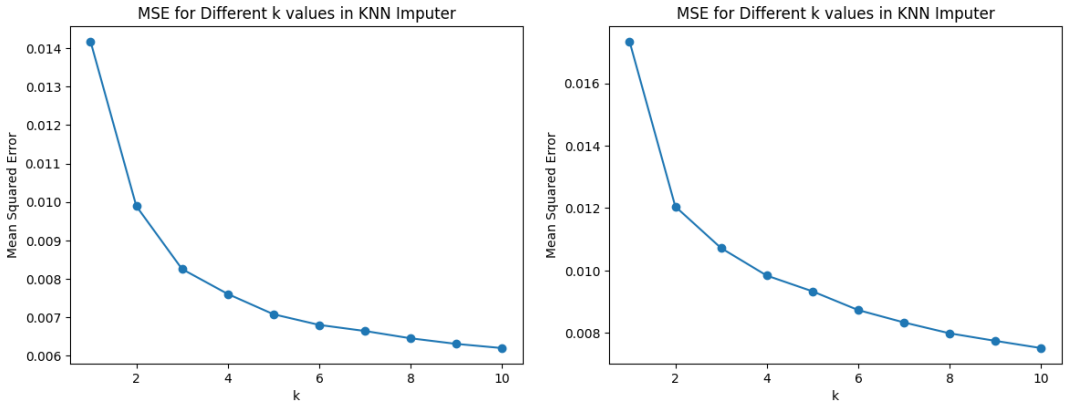
(a) Site F0010

(b) Site F0012



(c) Site F0014

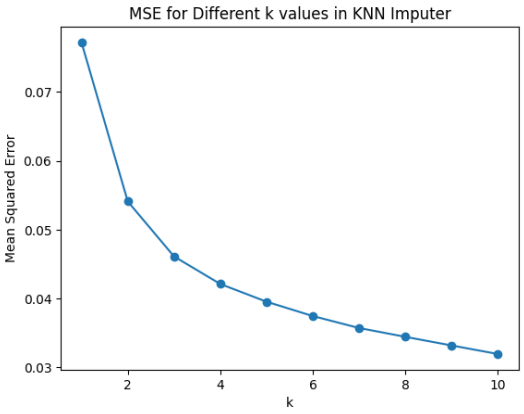
(d) Site F0016



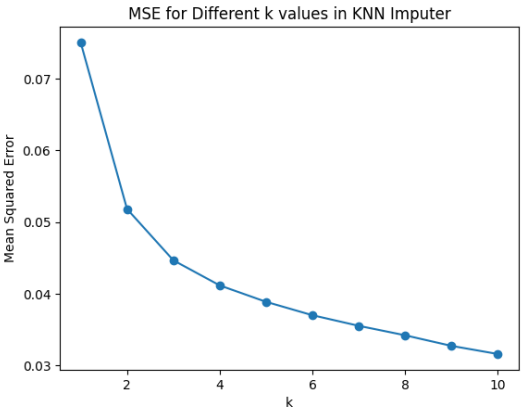
(e) Site F0022

(f) Site F0101

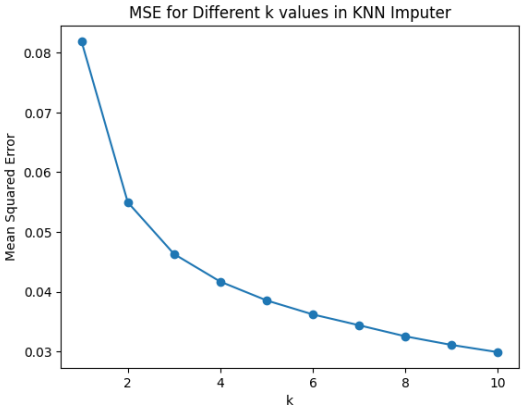
***Figure B.1*:** Plots showing the MSE vs k for all flow datasets.



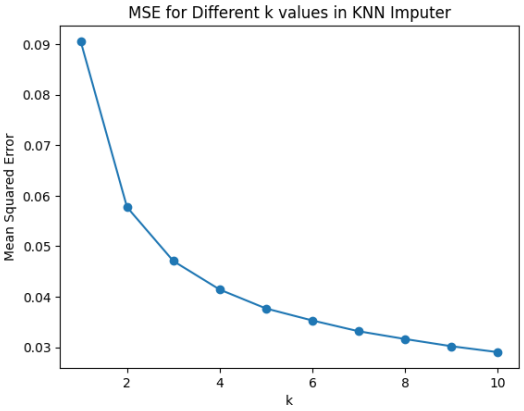
(a) Site S0010



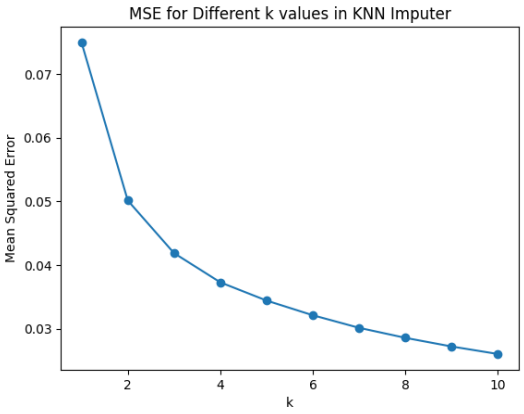
(b) Site S0014



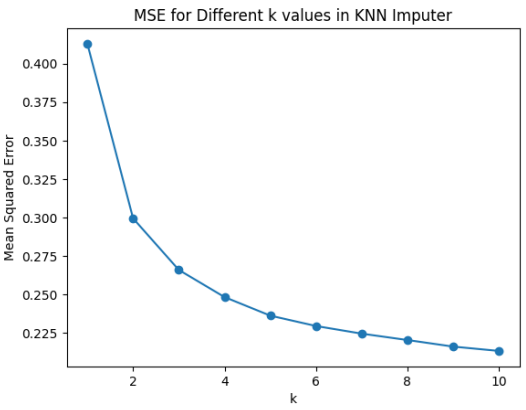
(c) Site S0015



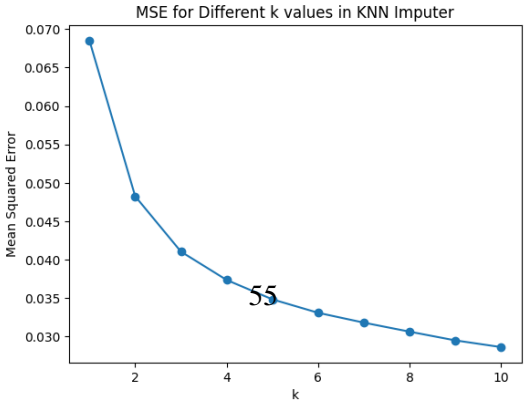
(d) Site S0016



(e) Site S0022



(f) Site S0024



(g) Site S0101

Appendix C

Github repo link

Link to the github repository containing all code files: `https://github.com/utkarsh11252/BradfordBeckAnalyses`