UNIVERSITY OF CAMBRIDGE
Department of Engineering

Master's Project Report

Quantification of Free Residual Chlorine for Water Purification

C-AJK61-3

Author: Neelay R. Sant
Supervisor: Dr. Alexandre Kabla

May 25, 2022

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed: ______________________________ *Date:* ____________________

Technical Abstract

Contents

1 Introduction and Background

Clean water is essential for human survival, and the topic of water purification is extremely important for ensuring a sustainable supply of clean water for everyone. The most prominent health concern associated with drinking water is contamination of it with pathogens that can cause deadly diseases. Due to its low cost, ease of use, and safety, chlorination is the most popular form of disinfection for drinking water. The main issue that arises from using this method is dosing samples appropriately [11]. Identifying an appropriate dose requires measuring the Free Residual Chlorine (FRC) of samples. Using a traditional in-line sensor (a sensor that could be in a pipe taking measurements) has limitations, with the main issue being their expense. This report evaluates the possibility of measuring other common water parameters in order to determine the FRC to an appropriate accuracy. This could then lead to developing a low-cost sensor to use in low-income areas.

1.1 The Importance of Water Purification

Access to clean drinking water is an extremely important and growing issue. 3 in 10 people lack access to clean drinking water worldwide [6]. Nearly half of the global population already live in areas of potentially water scarce areas for at least one month each year, which could rise to 5.7 billion by 2050 [26].

There are several processes involved in cleaning water from any source to produce drinking water. Contaminants can be in the form of particles, dissolved organics and inorganics, and microorganisms. Removing particles from the water can be done by filtration or sedimentation. Removing dissolved inorganic compounds can be done using ion-exchange or precipitation and dissolved organic compounds can be removed by adsorption. The most prominent contaminants in drinking water concerning health are microorganisms. There are 5-10 million water-related deaths worldwide each year, mainly from pathogens causing water-borne diseases, such as cholera and typhoid [6]. Pathogens in water are killed by disinfection.

1.2 Disinfection Methods

The main methods of disinfecting water supplies are ozonation, using UV light, using chloramines, and chlorination [9]. Ozone (O_3) is an unstable form of oxygen that decomposes very quickly to produce highly reactive free radicals. It has a higher oxidising power than even chlorine, and so is a very powerful and effective disinfectant [9]. It also does not leave behind any unpleasant odour or taste. However, the production of it is very expensive, requiring high amounts of energy and specialised equipment [21].

UV light can be used to disinfect water as it can break down the chemical bonds in DNA and proteins in microorganisms [21]. Its advantages are that it doesn't produce any byproducts, and it can limit potential regrowth. However, it also has a very high energy requirement, and its efficacy can be limited by more turbid water samples [15].

Chlorination is the process of disinfecting water samples by reacting them with chlorine. It kills microorganisms by oxidising bonds in their structure, and it is very effective at eliminating a wide range of pathogens due to its high oxidising power [18]. It is much cheaper to use than UV or ozone, and the dosing rate can be changed easily and flexibly. However, it can leave behind byproducts that can increase the toxicity of the water, such as trihalomethanes [9]. Also, the dose amount is of high importance, as too much can leave the water unpalatable, and too little can mean the water could still contain pathogens.

Chloramines can be used as an alternative and work in a similar way. They are produced by reacting free chlorine with ammonia and are also oxidising agents. They are more stable, so can provide long lasting residual disinfection. They also produce fewer byproducts [25]. However, due to their lower oxidising power, they are less effective than free chlorine at disinfecting samples.

Because it is the cheapest highly effective method, this report will only be concerned with chlorination as a disinfection method.

1.3 What is Free Residual Chlorine?

Chlorine can be added to water in two ways - as a gas (Cl_2), or in a hypochlorite (OCl^-) salt, such as sodium hypochlorite (NaOCl) [18]. The gas disproportionates in water into hydrochloric acid and hypochlorous acid (HOCl). HOCl and OCl^- are both strong oxidising agents and the sum of their concentrations in a water sample is the Free Residual Chlorine (FRC), but contaminants in untreated water affect the concentrations of these compounds as shown in figure 1.

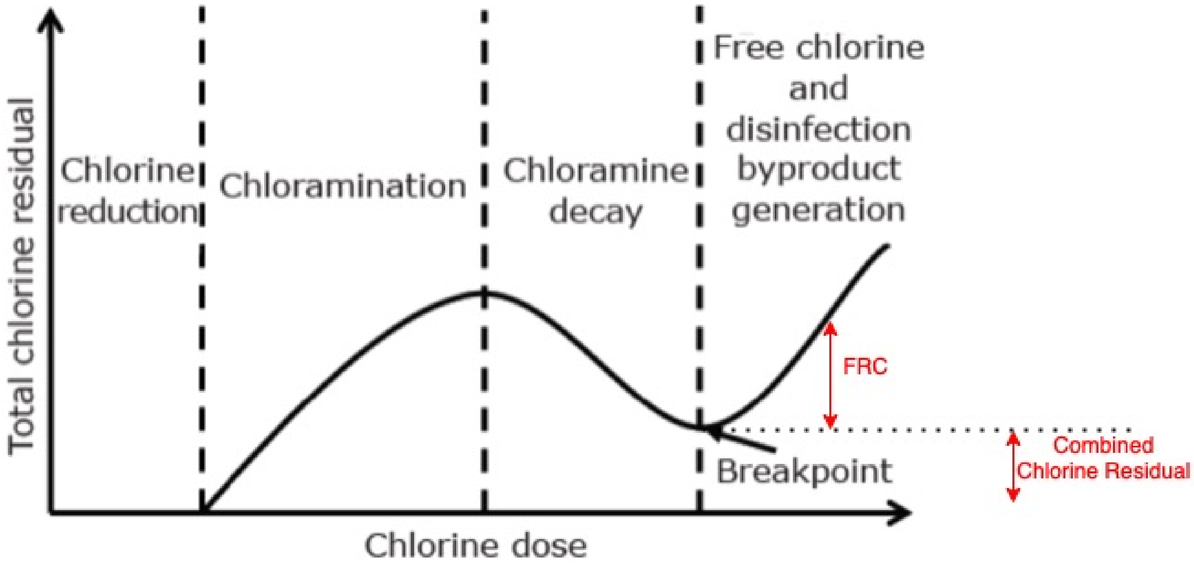


Figure 1: The Generation of Residual Chlorine in Untreated Water Samples [28]

In the reduction stage, chlorine will immediately react with reducing agents in the water, such as iron, manganese, and nitrites. No residual will form until these agents are completely oxidised. Once these are all reduced, chlorine reacts with nitrogen-based compounds (organic matter and ammonia), forming *Combined Chlorine Residuals*, such as chloramines. These residuals are weaker oxidising agents than FRC. As more chlorine is added, it will oxidise the chloramines further into compounds that are not oxidising agents, therefore reducing the residual. This occurs until the 'breakpoint' is reached, where we have a minimum amount of reducible chloramines in the sample. After this point, any chlorine added forms a proportional amount of FRC [28, 18].

The presence of FRC in a chlorinated water sample shows we are past the breakpoint, which means the sample is fully disinfected. The sample also then has a strong oxidising power, which is important as it can still disinfect the water more in case of further contamination. The oxidising power of OCl^- is around 1% that of HOCl, but both are much stronger oxidising agents than any chloramines [16]. Because of this, we can assume when water is safe to drink based on the FRC. The guidelines on how to treat a water sample based on the measured FRC (as set by the WHO [29]) are detailed in table 1.

FRC Range (mg/L)	**Short Label**	**Details**
$FRC < 0.2$	FRC too low	This is not high enough to be sure we are past the breaking point.
$0.2 \leq FRC < 0.5$	safe	This water is safe to drink immediately, but might not be safe to drink after storing, since it does not have much disinfection capability left.
$0.5 \leq FRC < 2$	safe to store	This water can be stored in a bucket or jerry can and should be safe to drink for 24 hours.
$2 \leq FRC$	FRC too high	This water can become unpotable due to unpleasant taste and odor from chlorine, and is above the WHO recommended level of FRC.

Table 1: A Table summarising the meaning, in terms of safety of consumption, behind the FRC of a sample according to the range in which the FRC fits. These are in accordance to WHO guidelines [29].

Fig 1 gives an indication of the difficulty of predicting the FRC in a sample indirectly. It shows how the FRC is not directly related to the chlorine added, and how different levels of contaminants can affect the amount of FRC produced by a specified dose.

1.4 Current FRC Measurement Methods

Reacting chlorine with water gives us three metrics to help us keep track of how much is needed; the dose is the amount of chlorine added; the demand is the amount that reacts with substances in the water through oxidation; and the FRC is the chlorine left uncombined. This gives us the simple relationship, $dose = demand + FRC$.

To maintain appropriate levels of FRC in water, the measurement of it is vital. This is so important because we must add enough to ensure complete disinfection, but increasing amounts make the water unpalatable and potentially toxic [1]. Figure 2 shows the distinction between different compounds present in water when chlorine is added. These groups can make the measurement of specifically FRC difficult, as they can all cause interference and affect all measurement methods.

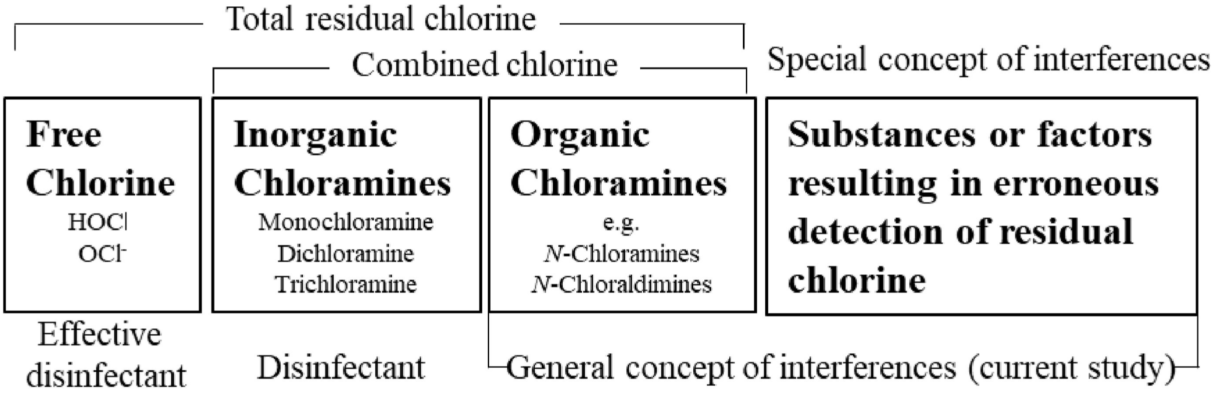


Figure 2: Forms of Chlorine and Interfering Compounds in Chlorinated Water [11]

The oldest method of FRC measurement is titration, reacting samples with starch iodide [14]. It is no longer commonly used, as it has disadvantages compared to modern instruments in accuracy, difficulty and measurement time. However, modern methods still derive from the principle of titration [11].

The electrochemical (EL) method is one commonly used for FRC measurement. Sensors using this method are often referred to as amperometric sensors. The method involves passing a small current between an anode and cathode. Due to chemical reactions involving FRC, there is a change in current which we can measure, that can then determine the FRC concentration [24]. It is based on the reduction of FRC at varying potentials and summarised by the following reactions:

in basic conditions:

$$OCl^- + 2H_2O + 2e^- \rightarrow Cl^- + 2OH^- \quad (1)$$

in acidic conditions:

$$HOCl + 2e^- \rightarrow Cl^- + OH^- \quad (2)$$

The EL method can also distinguish free chlorine from combined residual chlorine by controlling the pH and potassium iodide concentration in the solution. It is therefore suitable for measuring FRC in-line continuously [8]. Li, 2021 [11] states the determining range for the EL method is 0.01 to 1 mg/L. However, the company Sensorex sell an amperometric sensor that has a determining range of 0.01 to 10mg/L [5]. This leaves ambiguity in what range this method is suitable to be used in for determining FRC. In any case, the sensor is expensive (over $1000 [5]) and requires routine calibration against EPA-approved DPD colorimetric methods [13]. This is controversial within the water quality industry, as amperometric sensors have been shown to have large variance in the stated FRC for varying water parameters, such as pH [13].

The colorimetric-photometric method can be used for in-line as well as on-site analysis of FRC. It involves reacting chlorinated samples with one of two reagents - N, N-diethyl-p-phenylenediamine (DPD) or o-toluidine (OT). In this method, the FRC reacts with the colorimetric reagents, and the resulting change in colour is measured to identify the FRC concentration. Both reagents have determining ranges of greater than 4mg/L [11]. However, OT has been found to be potentially carcinogenic, so it is no longer commonly used as a reagent [22]. DPD is recognised to be used in the standard colorimetric method for FRC, but also has drawbacks in being sensitive to unexpected interference of other substances, such as high concentration of manganese or chromium [12]. It also has the drawback of requiring an added reagent for measurement that can't be mixed into the main supply.

For our purpose, we require a low-cost sensor that can measure in the range of 0-2.5 FRC at least. The expense and potential drawbacks of the methods mentioned have motivated this project to look at whether measuring other common water parameters are enough to serve as a proxy for measuring the FRC directly.

1.5 The Relation of FRC to Other Parameters

The FRC has potential to have strong relationships with other commonly measured water parameters. This project involves evaluating whether the measurements of these other parameters can be used to evaluate the FRC with enough accuracy. We expect the Oxidation-Reduction Potential (ORP) to hold the strongest direct relationship with FRC, due to the components of FRC being strong oxidising agents. ORP is a measure (in millivolts) of the tendency for a substance to oxidise or reduce others [3]. It measures of the electron activity in the sample, so depends on concentrations of all substances in the water, as well as temperature and pH. Chlorinated water samples are too complex to use the ORP to determine the FRC directly.

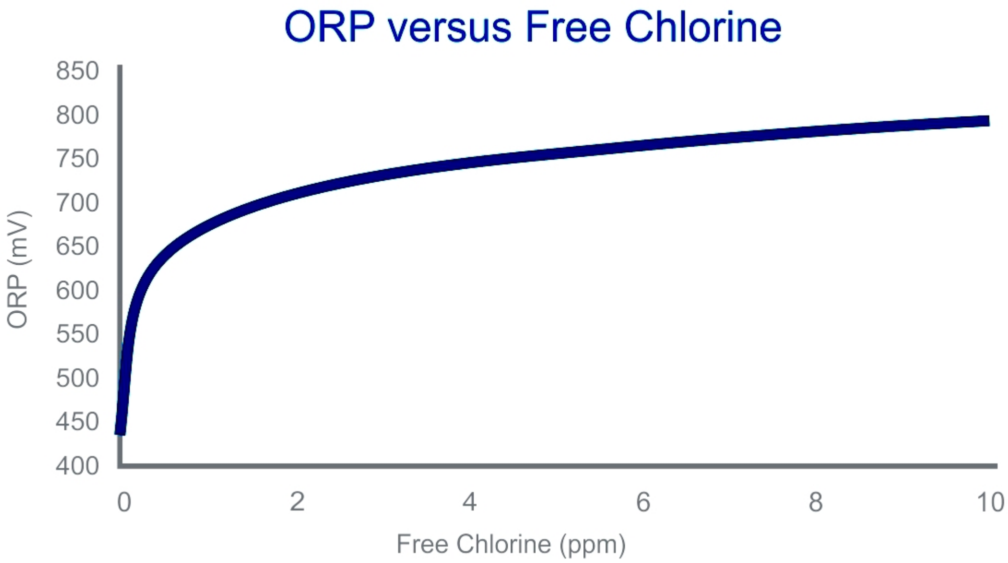


Figure 3: The Relationship between ORP and FRC [19]

However, it can also be considered as a measurement of the oxidising power of a substance. Since FRC has a high oxidising power, we can expect there to be a strong correlation between the FRC and ORP of samples, especially since the presence of FRC means the chlorine has already oxidised strong reducing agents in the water. It is like measuring how effectively it can disinfect a newly introduced contaminant, but not the amount of contaminant it can disinfect [20]. This relationship is signified in figure 3. The turning point in the graph shows the oxidising power of the samples is much less affected by more FRC after a certain threshold is reached. The location of this turning point, and how sharp the turning point is, are dependent on other substances other than FRC in the water.

Adding chlorine to water should not have much affect on the pH of the sample. However, the composition of FRC in the water will be affected by different pH. HOCl exists in water in an equilibrium with hypochlorite ions following eqn. 3. The equilibrium point is highly dependent on the pH of the sample. Acidic solutions will favour the HOCl form, and basic solutions will favour the OCL$^-$ form. Both are strong oxidising agents, but the oxidising power of OCl$^-$ is around 1% that of HOCl [28]. This means that the ORP of a specific amount of FRC would be significantly different at varying pH. Therefore, measuring the pH could prove useful in conjunction with measuring the ORP, as well as providing an indication of contamination level.

$$HOCl \rightleftharpoons H^+ + OCl^- \tag{3}$$

Conductivity is a measure of the capability of water to pass an electrical current. It is directly related to the concentration of ions present in the water, as well as their form, oxidation state and mobility [2]. Since chlorine introduces ions into the water when it is reacted with it, the amount added, and therefore the FRC, should have an affect on the measured conductivity of a sample. However, the conductivity will mainly be a good indicator of the level of contamination of a sample, which could be related to the chlorine demand, as high contamination would likely indicate high amount of reducing agents or pathogens in the water.

Another commonly measured water parameter is the Dissolved Oxygen (DO) level, although it is usually measured in the context of waters for aquatic life, as it is an essential component for those use cases. DO is not involved directly in reactions with FRC itself, but since it is an oxidising agent, it could feasibly also have an affect on the reducing agents in the water and the measured ORP in a sample. Therefore, it is possible that it will hold value when measured in conjunction with the ORP.

Turbidity is a measure of the relative clarity of water. It measures the amount of light scattered by water, and is an indicator of water quality as it is related to the amount of particles in the water [27]. Since FRC is soluble, it will not have a direct effect on the turbidity. However, turbidity can have a negative impact on the effectiveness of chlorine disinfection [10]. This could be because it reduces the ORP, so a fixed FRC could have a lower ORP in more turbid waters. Therefore, there may be benefit to measuring the two in conjunction with each other.

I have summarised the meanings and potential ways in which each parameter could be relevant into table 2. This is to aid quick reference to them later in this report.

Parameter	**Meaning**	**Relation to Chlorine**
Oxidation-Reduction Potential (ORP)	the oxidising strength of the water	high correlation due to the components of FRC being strong oxidising agents
Conductivity (Cd)	measures the capability of water to pass a current	chlorine forms ions in water so there should be a positive correlation. Also indicates the level of contamination in the water.
pH	$-log[H^+]$	controls the state of FRC in water, so should have an affect on the ORP measurement for a specified FRC.
Dissolved Oxygen (DO)	concentration of oxygen dissolved in the water	oxygen is also an oxidising agent, so having a high DO could change the ORP at a certain FRC as the oxidation power would be higher.
Turbidity (Tb)	how transparent the water is	negative impact on the effectiveness of chlorine disinfection, so may change the ORP at a specified FRC.

Table 2: Summary of Parameters and Their Possible Uses in Evaluating the FRC of a Water Sample

1.6 Project Aims

The aim for this project is to establish whether it is feasible to use the measurement of common water parameters to be able to quantify the FRC of drinking water samples that have been treated with chlorine, and to explore the accuracy that can be obtained. We will look at the direct relationships between each individual parameter and chlorine, as well as their accuracy and reproducability when being measured. This will give us an idea of which parameters are most likely to add information about the FRC when used in data analysis together.

We will then explore using different data analysis techniques to try and determine the FRC from the data. In the best case, we want to find a continuous mapping from our measured parameters to the FRC, i.e. we could output an estimated FRC from an input of the parameters we measure. In this project, this would be finding a direction We will evaluate methods trying to do this by calculating the cross correlation between the estimated FRC of the datapoints and their actual measured FRC.

Given the difficulty of producing a continuous mapping, we will also explore the possibility of forming a classifier on the datapoints, where each class will be a specified FRC range. The ranges we will classify points to will be based on those detailed in section 1.3, as this will tell us how to treat each water sample based on their class, i.e. the sample needs more chlorine, it is safe to drink, or the sample has too much chlorine. We will evaluate classifiers by their success rate in classifying datapoints.

2 Data Collection

2.1 Experimental Design for Data Collection

Samples from a range of water sources were dosed with varying amounts of chlorine, and measured for each of the included parameters, to build up a reliable dataset to be used in analysis. Each sample collected would vary in pH, conductivity, etc. due to different contamination levels before they were dosed with chlorine. This enabled us to look at the trends associated with changing FRC for a variety of different conditions.

2.1.1 Producing Accurate Chlorine Doses

The first step was to establish a robust methodology for creating and measuring samples, in order to give strong reproducibility of measurements. This involved establishing a method that produced chlorine doses accurate to what was expected, and finding the level of accuracy possible for that method.

For this step, samples of distilled water with varying FRC concentration were made using serial dilutions. Using distilled water for this stage made the process simpler and made the output more predictable, since the chlorine demand would be 0, so $dose = FRC$. Using a simple solution at first also meant readings from the other sensors were only being influenced by the one parameter being changed (the FRC), so their trends could be verified more easily.

The initial process involved making a concentrated ‘mother’ solution of 10mg/L by dissolving sodium dichloroisocyanurate (NADCC) tablets in distilled water, which would be diluted appropriately for creating samples of a specified dose. Samples were made with doses in the expected range of 0 to 2.5 mg/L, as this was the most relevant range for data analysis. The expected concentration was calculated from the proportion of mother solution added to plain distilled water. The resulting chlorine concentrations were measured using a colorimetric FRC sensor. For this, the sample was reacted with a reagent (DPD) to form a pink colour that could be measured.

Preliminary issues in the sample making process were that measured FRCs were consistently lower than expected, and had very high variation. This issue was either from the tablets used, the FRC sensor, or the mixing after dilutions not being as thorough as required. The sensor output was cross-checked with another sensor to ensure it wasn’t faulty.

The mixing was evaluated by varying the volume, the FRC concentration of the mother solution, and time spent mixing, to see which gave the best results. From this, we established the method of making samples to 100ml, using a 50mg/L mother solution, and shaking in glass bottles for a minute before and after the reaction time worked best, and gave FRCs as shown in figure 4. This method is detailed in figure 7.

The error bars indicate an error margin of 6%, which was the maximum percentage error associated with the diluting process and the FRC sensor. Figure 4 shows the procedure still gave a systematic error, with the mean 18.3% lower than expected, but the spread of values was more consistent. The average percentage error from this new mean was 7.95%, which was much better than previous techniques tried. The systematic error and remaining percentage error from the mean were associated with the FRC delivered by the tablets.

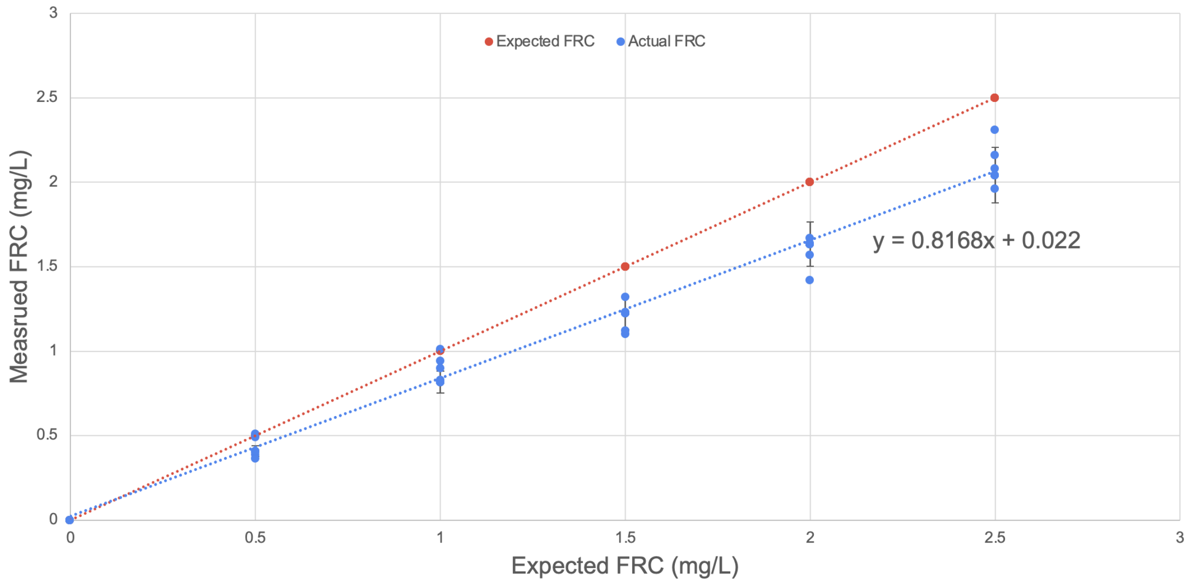


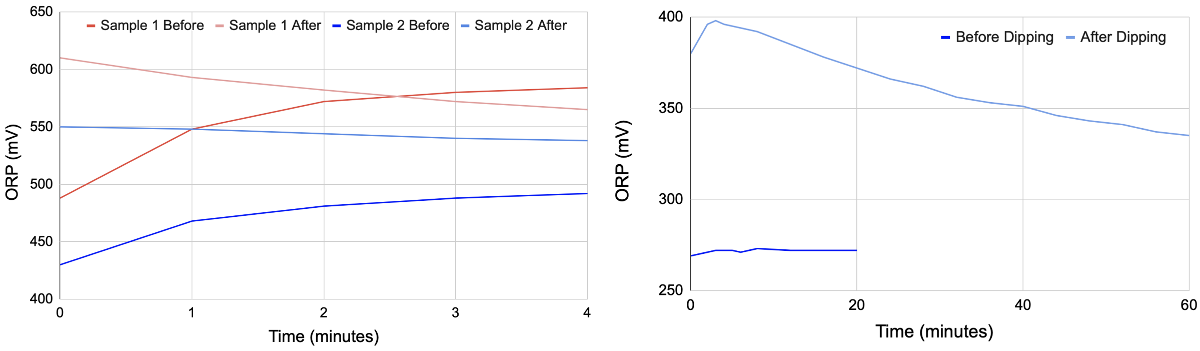
Figure 4: Expected vs Measured FRC for Distilled Water Samples

2.1.2 Verifying ORP Measurements

The outputs of the sensors used were verified by checking their consistency over time and seeing if samples of similar composition (similar FRC) and conditions (e.g. room temperature) would give similar readings. This verified the output of all sensors except ORP, which gave an interesting variation of results. The manual of the ORP probe stated it would take around 30 seconds to stabilise. In reality, it took much longer and the measurement varied depending on what the probe had been measuring beforehand. Figure 5a expresses these issues, as the time taken for the measurement to stabilise was much longer than expected, and had depended on the history of the sample.

To analyse this behaviour, the ORP probe was placed in some samples in quick succession to see the time taken to stabilise and the magnitude of influence on the reading. The results are shown in figure 5b. To start, the probe was placed in distilled water, when it had not been in any sample for more than 12 hours prior. This gave a reading of around 300mV, as expected. The probe was then dipped in a 10mg/L FRC solution for a couple of minutes. It was then cleaned and returned to measuring the DW sample again. Figure 5b shows the probe took around 10 minutes to stabilise at first, but took longer than an hour when it was heavily influenced by previous measurement. However, this was resolved with enough time between measurements, as it would return to taking 10 minutes after not being in use for some time.

To ensure the characteristics were not just due to a faulty probe, the output was compared to another ORP probe. Again both were placed in different samples in quick succession, with the results shown in Figure 6. We see that the second probe still exhibits the same odd patterns of being highly affected by recent measurements, and taking a significant amount of time to converge to a value. However, the convergence time is less, and the probe returns to having an unaffected measurement after 30 minutes, which isn't the case for the first probe. Because of this, all further samples were taken using the second probe. To mitigate the issues found, sets of samples were made in an increasing order of concentration, so that the ORP variation should be less. At least 30 minutes were allowed between ORP measurements so that the influence of history should be removed. These



(a) Readings from the ORP probe over time for two samples sample before and after measuring other samples. Note the different starting values and how the readings don't tend to the same value for either sample,

(b) Readings from the ORP Probe over time for a distilled water sample before and after dipping the probe in a concentrated FRC solution.

Figure 5: Graphs of Initial ORP Measurement Characteristics

changes were effective in giving results that matched samples of similar concentrations.

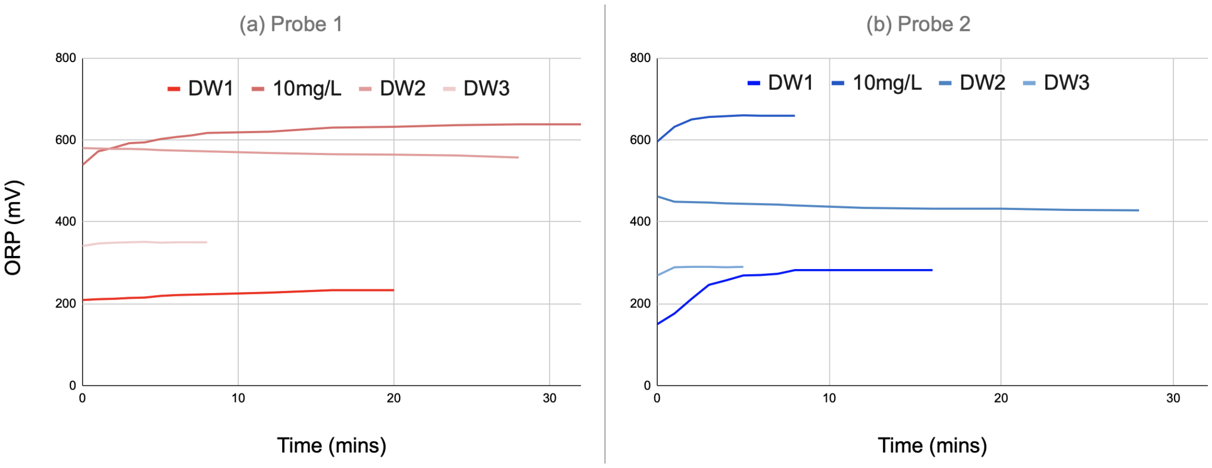


Figure 6: A Comparison of the Measurement Characteristics for 2 Different Probes in Different Samples. Both probes were initially placed in a distilled water sample (DW1). They were then placed in a 10mg/L FRC solution (10mg/L), followed by returning to the distilled water sample immediately (DW2). Finally, the probes were unused for 30 minutes and then measured the distilled water again (DW3).

2.1.3 Experimental Method

Once these issues were overcome, we established a process for making and measuring samples, detailed in figure 7. The high amount of shaking and long shaking periods were necessary as less than this still gave high variation in delivered chlorine doses. The thirty-minute reacting periods were because that is the time required for chlorine to react fully in water [29]. Making samples of similar doses together was also very crucial for being able to check consistencies and identify anomalies. This method was used for creating samples from all the sources detailed in table 3, except for some measurements taken in Kenya, which were from actual chlorine-dosed water dispensers.

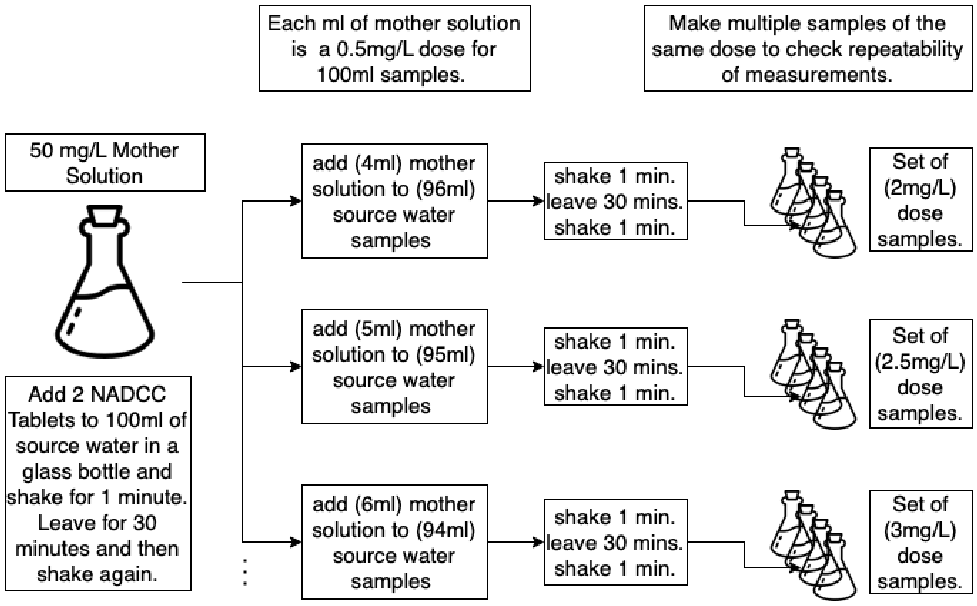

Each ml of mother solution is a 0.5mg/L dose for 100ml samples.
Make multiple samples of the same dose to check repeatability of measurements.
50 mg/L Mother Solution
Add 2 NADCC Tablets to 100ml of source water in a glass bottle and shake for 1 minute. Leave for 30 minutes and then shake again.
add (4ml) mother solution to (96ml) source water samples
add (5ml) mother solution to (95ml) source water samples
add (6ml) mother solution to (94ml) source water samples
shake 1 min. leave 30 mins. shake 1 min.
shake 1 min. leave 30 mins. shake 1 min.
shake 1 min. leave 30 mins. shake 1 min.
Set of (2mg/L) dose samples.
Set of (2.5mg/L) dose samples.
Set of (3mg/L) dose samples.

Figure 7: A Diagram showing the process of creating samples for one water source. The number of dose sets made would be enough to get a spread of FRC values in the 0 to 2.5 mg/L range (3 to 4 sets of 4 samples each). Numbers in brackets indicate these values would be changed in order to get FRCs in the range of 0 to 2.5.

Data Source	**Naming Convention**	**Details**
Distilled Water	Distilled	Distilled water from the laboratory.
CUED Tap Water	Tap	Tap water from the laboratory.
Midsummer Common 12/21	MC 1	Water from the river Cam at a place called Midsummer Common. It is downstream from tourist areas and close to animal habitats.
Midsummer Common 01/22	MC 2	This was again river Cam water from the same place but at a different time. This water required extremely high chlorine doses suggesting that it contained a high amount of reducing agent.
Jesus Lock	JL	This was another set of river Cam water collected upstream from Midsummer Common and closer to the tourist area.
College Pond	Pond	This was water collected from a college pond. The pond hosted a variety of aquatic life.
Swimming Pool	Pool	This was water collected from a Swimming Pool.
Ivonangya, Kenya: 09/21	Kenya Ivonangya 1	Blue Tap were able to collect water measurements in Kenya on two separate occasions in two places. This set was from the first occasion.
Ivonangya, Kenya: 03/22	Kenya Ivonangya 2	This was water collected in Kenya on the second occasion.
Mumo, Kenya: 09/21	Kenya Mumo 1	This was water collected in a different location in the first trip to Kenya.
Mumo, Kenya: 03/22	Kenya Mumo 2	This was water collected in the second trip to Kenya.
Ivonangya, Kenya Jerry Cans	Kenya jerry cans	This was water collected from Jerry Cans filled from the water supply in Ivonangya.
Kyuso, Kenya: Rock Catchment	Kenya Kyuso	This was water from a rock catchment in Kyuso, Kenya
Hostel in Kenya	Kenya Hostel	This was water from a rainwater collector in a Hostel in Kenya.

Table 3: Summary of Water Sources Included in the Dataset with Naming Conventions used in the Report Graphs and Details about the Source

2.2 Data Collection Results and Analysis

2.2.1 Conductivity

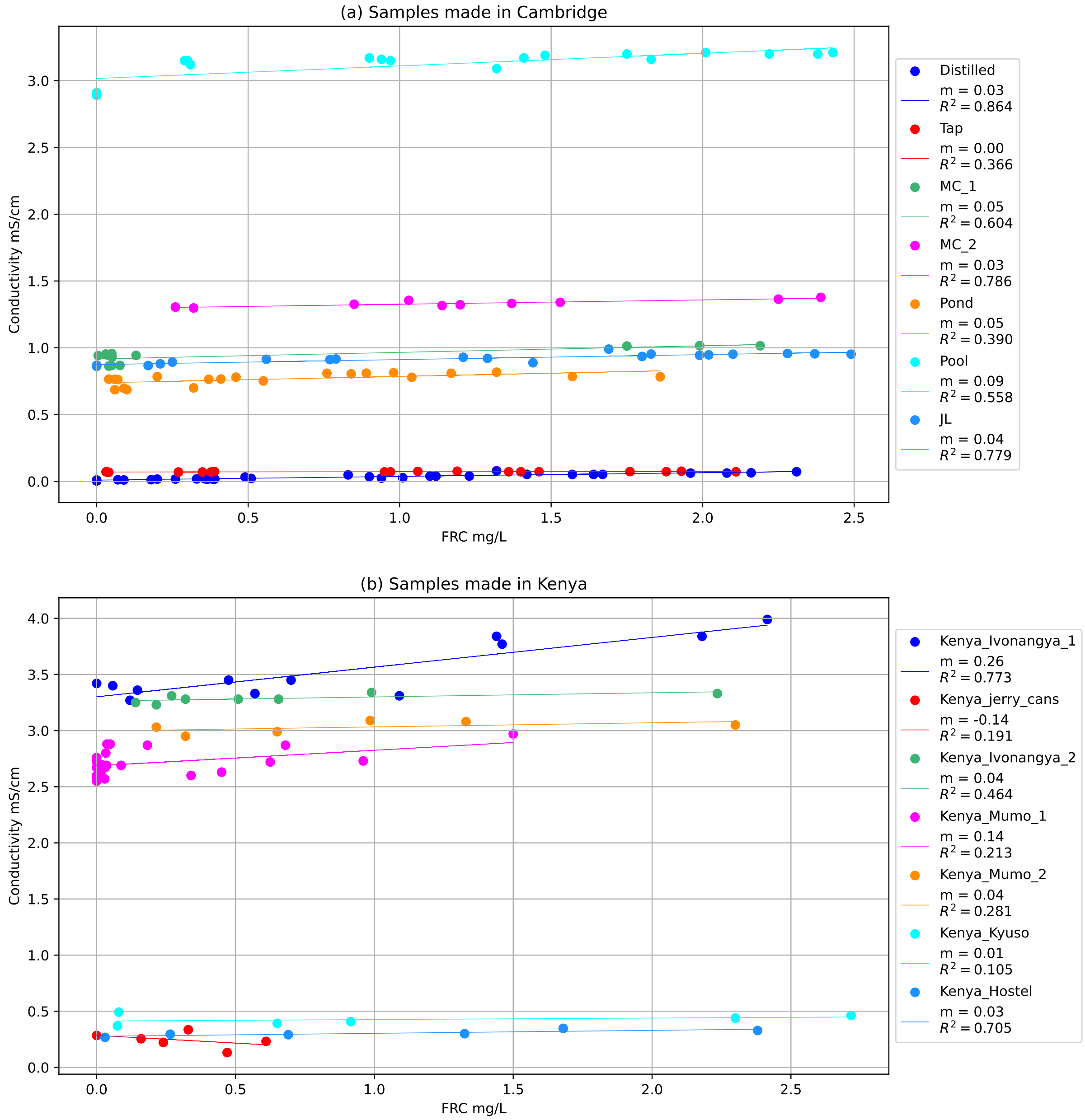


Figure 8: Conductivity Trends for Different Sets of Samples, with gradients (m) and determination coefficients (R^2) shown in the legend. Naming conventions of sets are detailed in table 3.

Figure 8 shows that across the different sources, there are commonly slightly positive gradients, with varying values of determination coefficients. This is in line with what we expected from table 2. The distilled water dataset has a strong determination coefficient. This can be expected due to the simplicity of the makeup of the water. The only change being made is additional ions in the water, so it makes sense that the correlation is strong. The variation in this relationship across all samples is most likely due to the complexity of these different waters causing more interference when the ions are added. Whilst conductivity is positively correlated with inorganic dissolved solids, organic compounds have poor conductivity, and could cause enough interference to mitigate the influence of increased FRC. Tap water can contain contaminants such as nitrates that might have this effect [7]. The takeaway from this graph is that conductivity does have an identifiable relationship with the FRC, but it is very source dependent, i.e. in a model we would need more information on the source to be able to use the conductivity to deduce the FRC.

2.2.2 pH

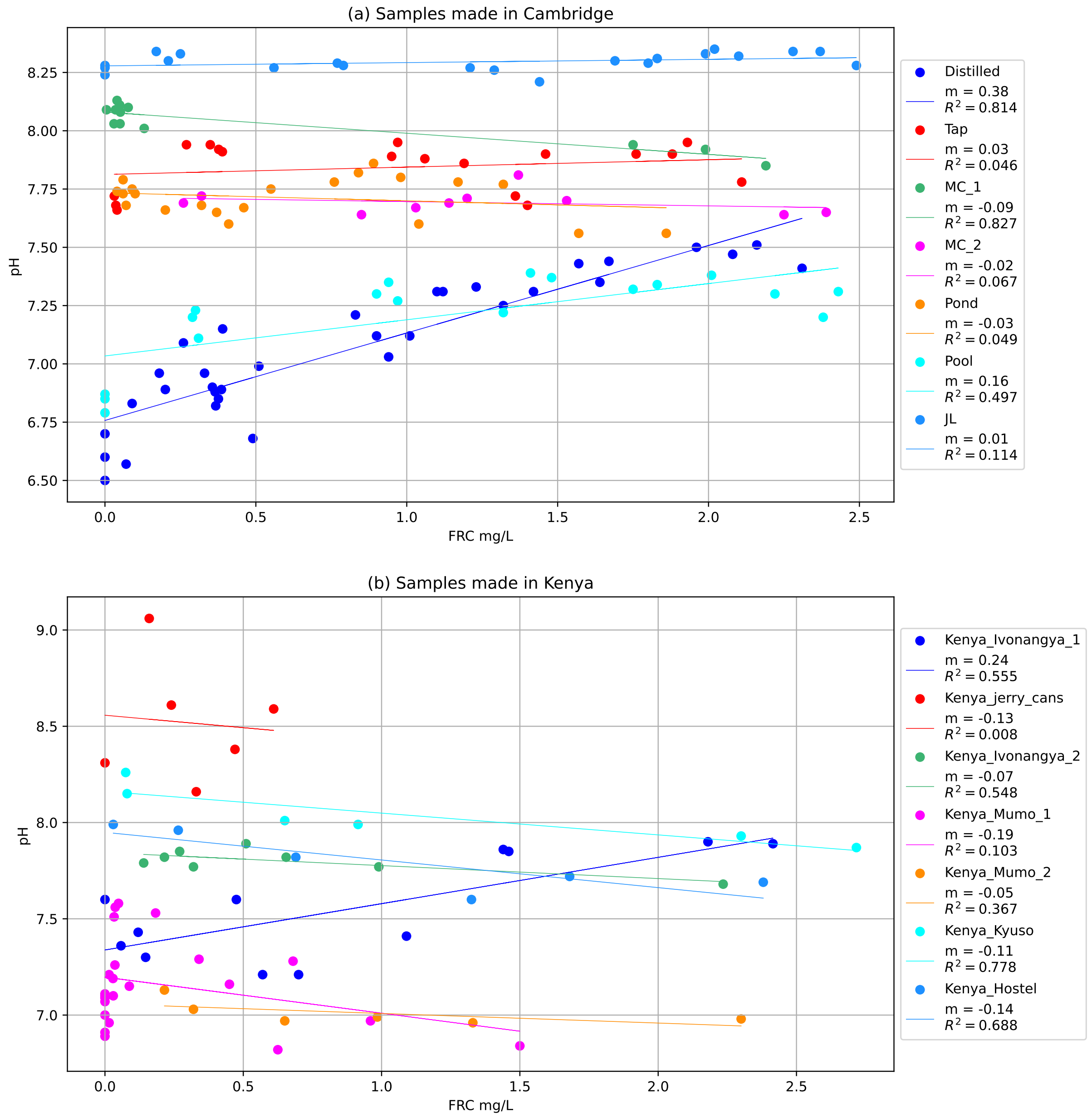


Figure 9: pH Trends for Different Sets of Samples, with gradients (m) and determination coefficients (R^2) shown in the legend. Naming conventions of sets are detailed in table 3.

The components of FRC exist in an equilibrium in water (shown in eq. 3) that is highly dependent on pH. Higher pH levels (over pH 8) cause the equilibrium to lie further to the right. From fig. 9, we see varying correlations in our data. Distilled water has a convincing positive correlation with a high R^2 value, but many sample datasets have much lower or negative gradients with varying R^2 values. It is therefore difficult to conclude whether we will get any information directly on the FRC from the pH measurement. However, from table 2, we know that the ORP will likely have dependency on the pH when being used to predict an FRC, due to the equilibrium of FRC.

2.2.3 ORP

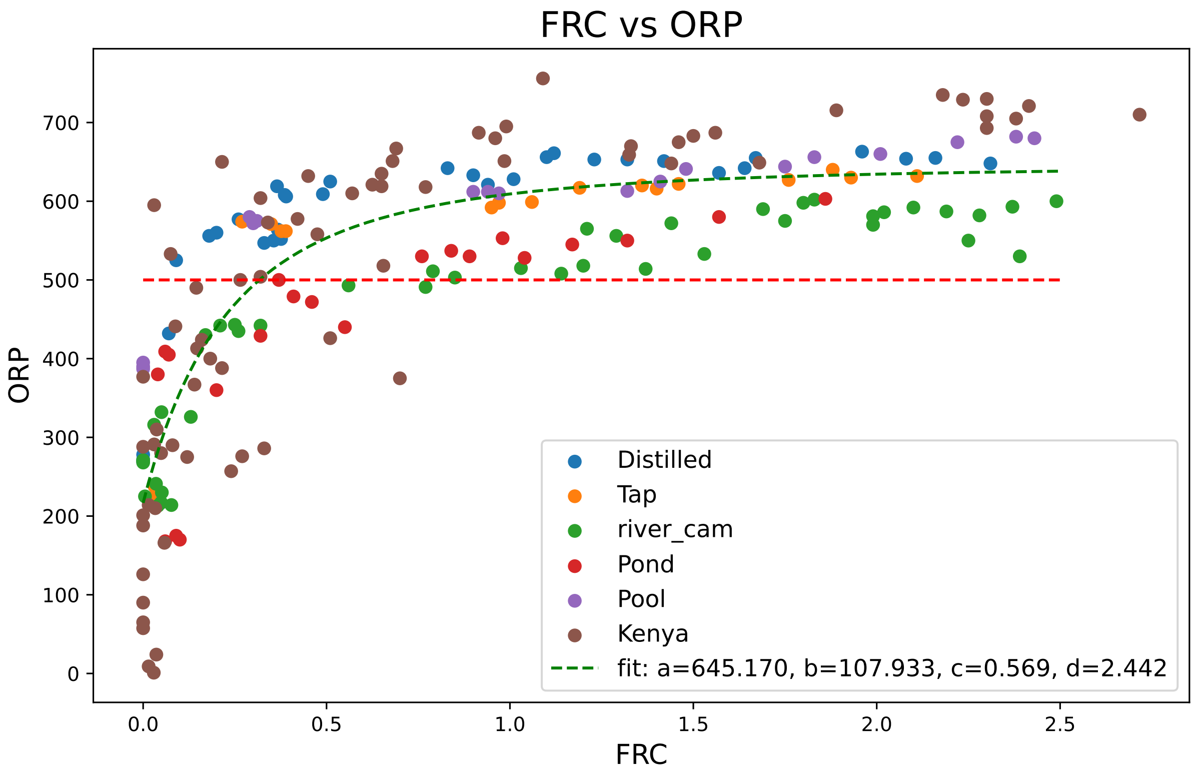


Figure 10: ORP Trends over Sets of Samples with a Plotted Trend Line. The coefficients given in the legend follow those detailed in eqn. 4.

Since HOCl and OCl$^-$ are both strong oxidants, we expect ORP to have a correlation with our FRC. This is confirmed in figure 10. Although there is slight separation between datasets, they all follow a general trend that seems to be fairly independent of the source. Since we want to be able to predict our FRC from the ORP, a trend line has been fitted to the data of the form in eqn 4 as an estimate of our true relationship. This has an RMS error of 87.7mV. The inverted relationship is given in eqn 5, which shows the mapping from ORP to FRC. We can see the graph shows a plateau for $ORP > 500$ (indicated by the red dashed cutoff line), with a steep rise before that. This is a very good sign of the potential of this measurement to predict our FRC.

$$y = a - \frac{b}{(x+c)^d} \quad (4) \qquad\qquad x = \left(\frac{b}{a-y}\right)^{\frac{1}{d}} - c \quad (5)$$

However, the challenges faced using the ORP sensor suggest it is a less reliable measurement than our other variables, as it is more difficult to reproduce, and getting a dependable reading in the field may prove more difficult. Also, for a specific FRC, ORP can vary for two reasons [17]. The first is from OCl$^-$ having a lower oxidation power than HOCl, If FRC is made up more of OCl$^-$, then the ORP will give a lower reading. This can occur at higher pH values as described before, so them being used in conjunction might yield more information. The second reason is that other oxidants, such as DO, and reductants, such as sodium, will also influence the ORP. Again this could mean that using DO and conductivity measurements in conjunction with ORP could give us a stronger model.

2.2.4 Turbidity

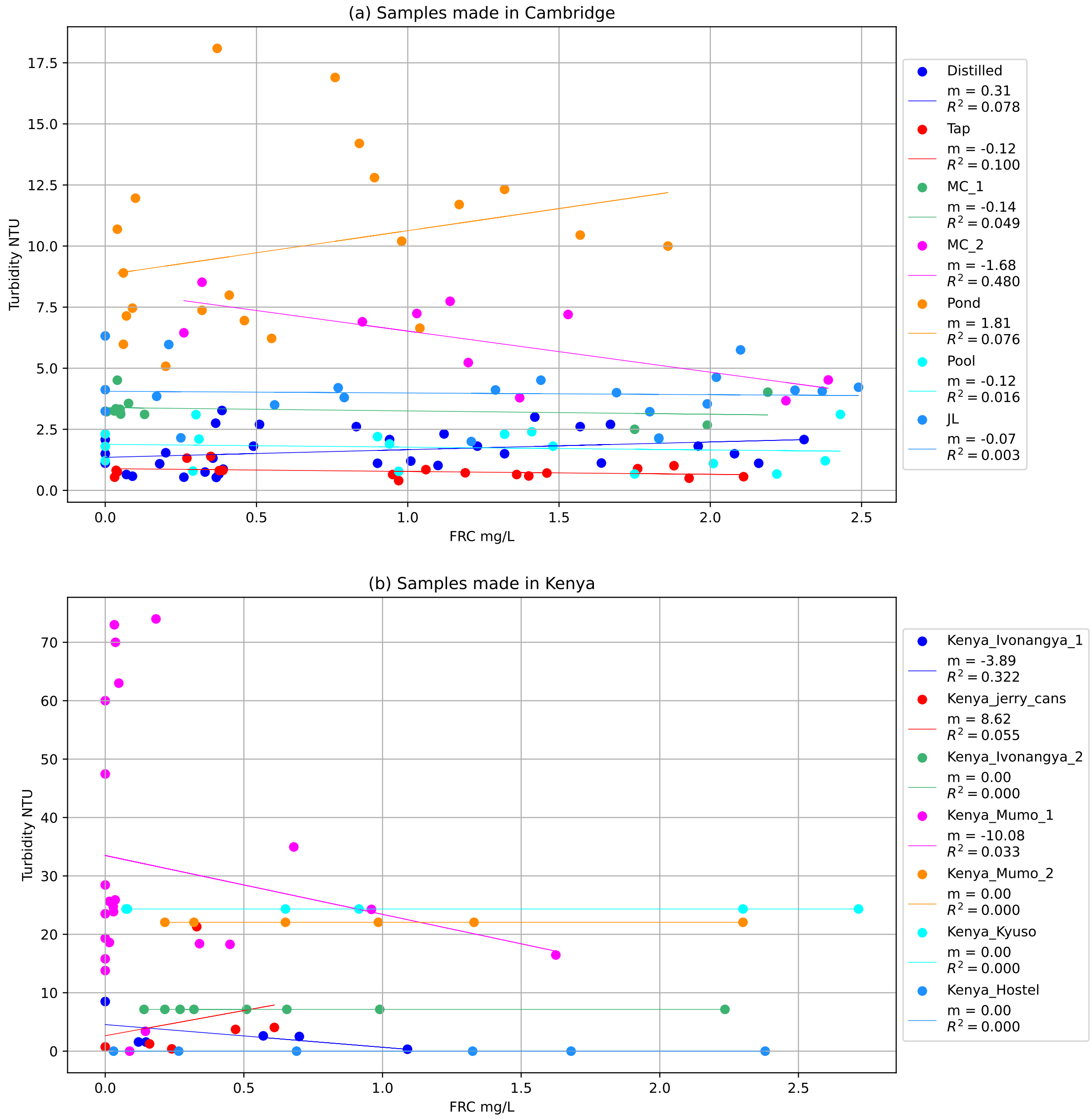


Figure 11: Turbidity Trends for Different Sets of Samples, with gradients (m) and determination coefficients (R^2) shown in the legend. Naming conventions of sets are detailed in table 3. (NTU = Nephelometric Turbidity Units)

Fig. 11 shows all sets have very low R^2 values for their linear relationships between

FRC and turbidity, indicating little to no correlation. Since the levels of chlorine we look at all fully dissolve in water (so they don't affect how transparent the water is), it makes sense we don't see much of a trend. Turbidity is really a measure of particulates in the water, which are usually removed by filtration. Overall, this graph shows that we do not expect to see direct information given on the FRC from the turbidity measurement alone. However, the reasons detailed in table 2 about turbidity affecting the effectiveness of chlorine mean we could still see some use in its measurement when used in conjunction with ORP.

2.2.5 Dissolved Oxygen

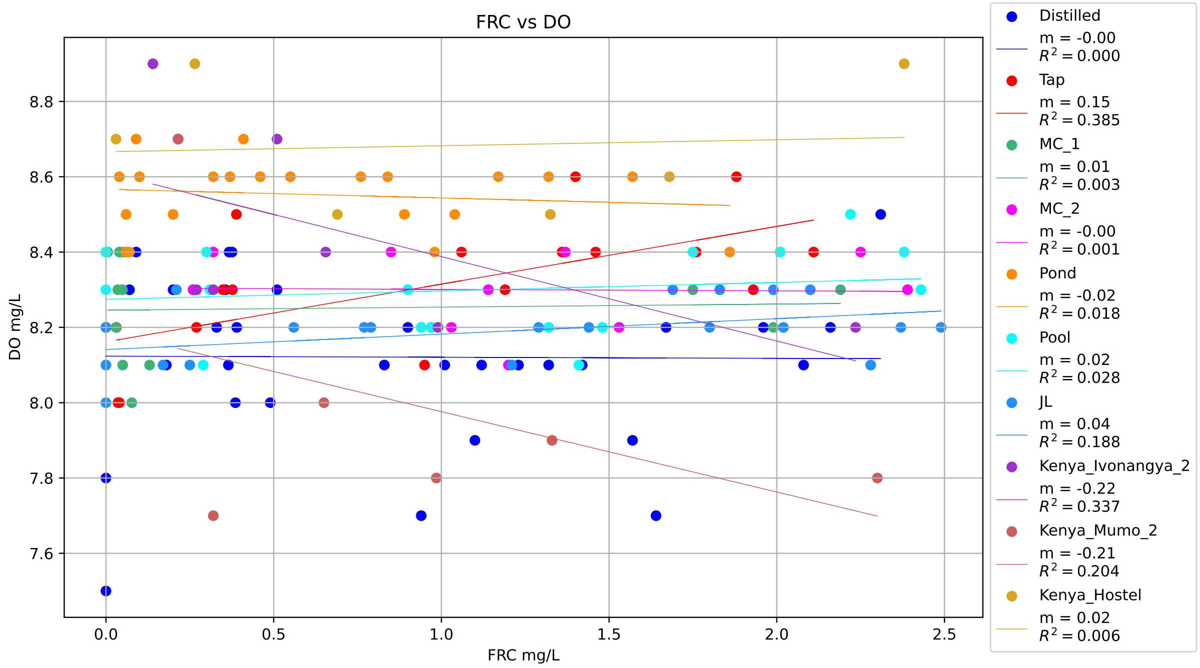


Figure 12: Dissolved Oxygen (DO) Trends for Different Sets of Samples, with gradients (m) and determination coefficients (R^2) shown in the legend. Naming conventions of sets are detailed in table 3.

Fig. 12 shows that no sets hold a strong correlation between FRC and DO levels, which is expected as they are not involved directly in reactions together. We see that the range over which the dissolved oxygen varies is also quite small (most datapoints are in the 8 $\rightarrow$ 8.6 range, which is quite tight. Waters analysed for aquatic life can appear in the range of 3 to 12 mg/L [4]. This suggests that the water samples that we will be using will not vary much in DO levels and that this is representative of waters used for drinking water in general. Therefore, it is likely it will not be a useful measurement to determine FRC. High DO levels will push our ORP curve higher as oxygen is also an oxidant, so the measurement could have been relevant; the small variance that we get in these waters are probably not enough for this effect to have an impact.

2.3 Data Collection Conclusions

From the analysis of individual parameters against the changing FRC, we see the ORP provides the most valuable information, as it holds a relationship with FRC that is reasonably independent of the source. The conductivity held some relationship with the FRC that was somewhat consistent over the different sources, but the relationship was very source dependent. The pH did not seem to have any direct correlation, but we still expect it to be important when used in conjunction with the ORP.

The turbidity measurement had a very weak correlation with the FRC. Due to possibly affecting the effectiveness of chlorine, we may see some use of measuring it in conjunction with the ORP, but otherwise it will likely not be a useful measurement to include.

The DO had a vary small range of variation over all the samples, and also did not seem to have any correlation with the FRC. Because of this, we think it is likely that it will not be useful in determining the FRC.

3 Data Analysis Approaches

The aim of the data analysis is to be able to determine the FRC from a range of different measured variables. In the best case, it will produce a mapping of the variables to one continuous dimension that is aligned with the FRC. This is a dimensionality reduction problem, as we will be starting with the number of dimensions equal to the number of measured parameters, and want to reduce that to one dimension. If this is infeasible, we want to be able to classify datapoints into groups defined by ranges of FRC. This is relaxing the constraint of having a continuous dimension, but still allows us to mark datapoints as being in a certain FRC range.

For the classification, we will first divide the FRC range into groups detailed in table 4. These are based on the WHO guidelines mentioned in table 1, but we have combined the $(0 \rightarrow 0.2)$ and $(0.2 \rightarrow 0.5)$ groups into one group. This is because it is still reasonable to treat these samples as having too little, enough, and too much chlorine, and also gives us better separation between the classes. Distinction between the two groups that we have combined proves very difficult in analysis, so we decided that the improved simplicity of classification was enough to justify combining the groups.

Class	**FRC Range (mg/L)**	**Short Label**	**Details**
1	$FRC < 0.5$	too low	should be dosed with more chlorine
2	$0.5 \leq FRC < 2$	safe	safe to drink and store for 24 hours
3	$2 \leq FRC$	too high	water unpotable

Table 4: Classification Ranges for Data Analysis classes. This method will be referred to as the 3-Class Classification.

We will also look at conducting a classification purely on identifying whether the sample is below or above 0.5 FRC. This will only be able to serve as a marker on whether the sample is fully disinfected or not. We will refer to this method as the '2-Class Classification'. This distinction is still useful in the field as a consumer could use the classifier to be happy the sample is disinfected, and judge the potability based on smell and taking a sip.

Since this analysis is on a low number of datapoints, it is infeasible to look at deep learning models for analysing the data. For this problem, This report looks at using Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). All these methods work on using Singular Value Decomposition (SVD) on matrices produced from the data, which is feasible for a small dataset. The effectiveness of each decomposition is evaluated by:

1. Finding the correlation between the datapoints mapped onto eigenvectors found and their FRC. This will be a score between 0 and 1 from the magnitude of the correlation.

2. Making 3-class and 2-class classifiers based on the decomposition from each method, and calculating accuracy scores on the classification (number of points correctly classified).

Both of these scores will be calculated based on training/test data splits, with a random 80% of the datapoints being used in the training set, and the remaining 20% being used in the test set to calculate correlation or classification scores. These splits and calculations will be repeated 50 times, and the average scores over all the runs will be reported. This is in attempt to not have our results be based on over-fitting, and give a reflection on the scores we would obtain seeing new data.

This method has a flaw in that in the real world, we wouldn't have training data on samples from a new water source we are evaluating. Therefore, we will also look at making our training/test splits based on including and excluding full sets of data from a particular source (e.g. our test set will be all samples from Ivonangya, Kenya). We will do this as an evaluation on our best performing analysis methods to compare them against each other.

The decision on what would count as a 'successful score' for correlation or classification is not concrete and quite arbitrary. This is because the final dataset we have to work with is quite small (less than 200 datapoints), so we are unlikely to get scores high enough to be acceptable to use in the field (less than 1% error perhaps). In this case, we want scores high enough to say we have a strong correlation between the method output and FRC, enough to justify taking the analysis forward and optimising it in the future (e.g. using a much larger dataset). Therefore, we would suggest a correlation of greater than 0.8, or a classification accuracy higher than 0.9.

4 ORP as a baseline

The measurement of the ORP holds significant correlation to the FRC by itself; the correlation (COR) over all the datapoints of measured ORP and FRC was 0.704. This suggests using the ORP alone could give us a reasonably accurate FRC, or be used to classify points. We can evaluate this using the trend fit of the ORP calculated (eqn 5) to give predicted FRC values that we can correlate to actual FRC values. It can also give us ORP bounds for each class, based on the ORP values that map to the FRC bounds of each class.

For this analysis, the training/test splits will give training data that we will use to find values for the parameters in eqn 5, as well as ORP bounds for classification in the three class and two class case. We will then use the test data to calculate correlation and classification scores from the trend fit and ORP bounds calculated. We will repeat this 50 times and get average scores, which we can use to compare against other data analysis methods. The results are shown in table 5. One point to note is that because our curve has a plateau, ORP values above the plateau will not map to an FRC. Because of this, we will assume that these points have > 2.5 FRC, and so will be classified as such.

	COR	**3-Class**	**2-Class**
ORP Method	0.527	0.470	0.470

Table 5: ORP Correlation (COR) and Classification Scores. Note '3-class' refers to the classification score with three classes, and '2-class' is the score with two classes.

Table 5 shows that ORP does not do a good-enough job by itself. The reduced correlation when compared to the full dataset's correlation (0.704) is expected. We are using less data in each run to calculate the curve fit, and we are calculating the correlation on previously unseen datapoints, compared to seen datapoints for the full dataset value reported earlier. Interestingly, the classification score is identical for the 3-class and 2-class cases. This likely means that all mistakes made by the classifier were in the distinction between datapoints above or below 0.5 FRC. From figure 10, we expect the misclassifications to be coming from the datapoints with high ORP, low FRC or datapoints with lower ORPs at high FRCs. These scores would likely improve with a larger dataset for calculating the best fit parameters and scores.

5 Principal Component Analysis (PCA)

5.1 Background

PCA is an unsupervised learning technique that aims to reduce the dimensionality of a dataset whilst minimising information loss. It does so by finding eigenvectors, in the space of the variables measured, that maximise the variance between datapoints. This is in order to maximise the separation between datapoints, so that it is easier to identify the class they belong to. These eigenvectors are called 'principal components (PCs)'. The first step is normalising the data so that the dataset has a mean of 0 and unit standard deviation for each variable. The first principal component (PC1) is then found to be the direction in the space that, when all datapoints are mapped on to it, maximises the sum of the magnitudes. This is equivalent to finding a line of best fit over all the data in the normalised space. The second principal component (PC2) then does the same but over all directions perpendicular to the first. This repeats for as many principal components as required.

5.2 Methods of Implementation

The motivation behind using PCA for finding a dimension closely aligned to the FRC is that the changing FRC would be a major cause of variance between samples, so a principal component might be encouraged to align with it. From the data collection results, it appeared that the ORP, pH and conductivity (Cd) were the best parameters to begin with in order to use in analysis, as they appeared to deliver the most information on the FRC. After this, we looked at including the dissolved oxygen (DO) and turbidity (Tb), to see if they improved the results gained.

PCA does not have any classification method associated with it directly. It is mainly used for dimensionality reduction. However, for the purpose of comparing the different analysis techniques explored, it is feasible to implement an ad-hoc method of classifying our datapoints after finding a PC. We will do this by labelling our training datapoints with the class they belong to, and calculating a mean for each class. We can then project this mean on to the PC found. We then classify test datapoints by projecting them on to the PC, and assigning them to the class whose mean is closest to the datapoint. This method is assuming identical variances in each class, which we know not to be the case as our FRC ranges are not equal in size. However, since PCA does not have any attributed classification method, it can serve as a simple approach that we can compare to in-built classification methods used in our other analysis techniques.

PCA can only involve linear mappings from variables to potential principal components. Figure 10 shows that the ORP definitely does not have a linear relationship with the FRC. However, it is feasible to consider dividing the ORP trend into two 'linear' regions. One for the steep trend at the start, and one for the shallow trend for greater FRC. PCA could then be used to find a mapping for the separate regions, using the cutoff line marked in figure 10.

The idea behind this would be that a sensor in the field could use the initial ORP measurement to separate datapoints into a 'potentially still infected' group and a 'potentially too high FRC' group, based on if the ORP was below or above the cutoff. The mapping provided by the PCA could then identify if the estimated FRC was an issue (i.e. under 0.2mg/L or over 2mg/L). For this analysis, we used correlations of each region's principal component with the FRC, and then a classifier score on marking samples 'problematic' (too low or too high FRC) or 'unproblematic' (safe FRC).

An important note to make is that even if the FRC does not align with PC1, it could still be useful if it aligns with a subsequent one. This would mean that the FRC isn't the main cause of variance among samples, but the prior principal components could be accounting for variance between sources, that we anyway want to be removed. For example, if two sources have vastly different 'baseline' conductivities, before any chlorine is added, but hold a similar trend for change in conductivity with changing FRC, then PC1 may align with the conductivity between sources, and PC2 could align with the conductivity change due to FRC.

5.3 Results and Discussion

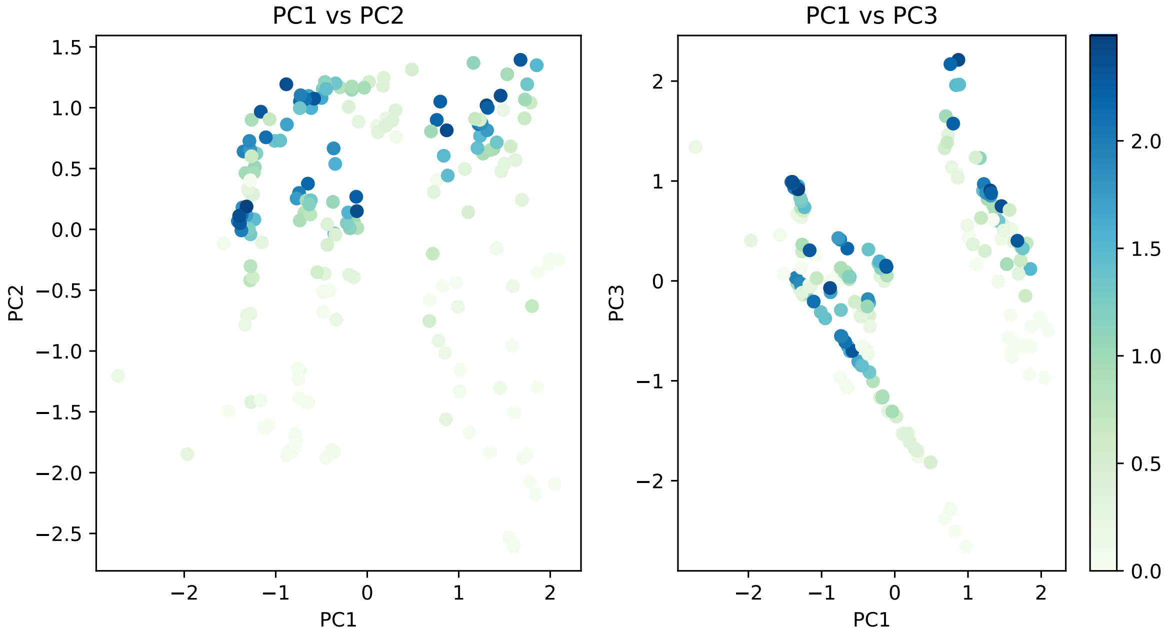


Figure 13: PCA results on analysis using the parameters Conductivity, pH and ORP.

	PC1	PC2	PC3
COR	-0.193	0.610	0.355
Vector	[0.72 -0.69 -0.12]	[-0.12 -0.29 0.95]	0.69 0.66 0.29]

Table 6: Initial PCA Correlations between PCs found and the FRC of All Datapoints, within a 0 to 2.5 FRC range and including DO, Tb and Cd parameters.

Figure 13 shows the datapoints plotted in the space of the three PCs found using the ORP, conductivity and pH parameters. The shading indicates the FRC, as shown on the right. We can see that out of the three PCs, PC2 appears to be the best aligned with changing FRC. This is supported by the largest magnitude correlation being associated with PC2. However, although a strong correlation, it is not too close to a direct relationship, and still has a lot of error within it. For the plot of PC1 vs PC3, we see that there is a distinct separation between two sets of datapoints. Both PC1 and PC3 have a high weight on the conductivity, as seen in table 6, by the high magnitude of the conductivity component of both of their vectors. Therefore, we can expect the conductivity to be the cause of the separation. We have no datapoints in the 1.5 to 2.5 mS/cm range, as seen in figure 8, which is highlighted in this plot. Unfortunately, this separation is not helpful in identifying the FRC.

Trial	**Parameters**	**Weightings of best PC**	**|COR|**	**3-Class**	**2-Class**
1	Cd, pH, ORP	[0.18 0.26 0.92]	0.603	0.478	0.786
2	Cd, pH, DO, ORP	[0.33 0.44 0.37 0.62]	0.536	0.394	0.730
3	Cd, pH, Tb, ORP	[0.49 0.57 0.27 0.57]	0.475	0.372	0.715

Table 7: PCA Correlation (COR) and Classification Scores for Different Settings. The ‘best PC’ refers to the PC that aligns best with the FRC. The ‘weightings’ are the average magnitude (over the 50 runs) placed on each measured parameter in the best PC. They correspond to the parameters in the same order that the parameters have been listed in the 'Parameters' column. i.e. in ‘Trial 1’, the PC that aligned best with the FRC, on average, had a 0.92 weighting on ORP. Note ‘3-class’ refers to the classification score with three classes, and ‘2-class’ is the score with two classes. Both classifications have been done using the ad-hoc method detailed in section 5.2.

Table 7 summarises the average scores obtained for different settings of PCA conducted. In trial 1, we see that the ORP ended up being very heavily weighted in the PC that was best aligned with the FRC. This is expected, since the two parameters are strongly correlated. Both the pH and conductivity are given low weights, which means the variance in them was largely not associated with the FRC. This was expected from the results in the data collection. We see that the correlation and 3-class classification scores are slightly improved from the ORP along, although not very significantly.

However, even though we used a simple ad-hoc method, the 2-class classification score has dramatically improved. This could be because we are using a linear relationship rather than the non-linear relationship used for the ORP alone. If we imagine a linear LOBF plotted over figure 10, the new ORP cutoff for 0.5 FRC would be less. This would make more misclassifications on (high ORP, low FRC) datapoints, but fewer misclassifications on (low ORP, high FRC) datapoints.

In trial 2, we see that incorporating the dissolved oxygen measurement had a negative effect on all our scores. This means that the measurement included much more variance that wasn’t associated with the FRC. PCA is just trying to find directions of maximum variance, and we are using it in the hope that the variance is being caused by the FRC, as it is a parameter we have changed throughout the dataset. Unfortunately, the DO seems to vary much more in ways not due to the FRC, so it makes our analysis output worse. From the reasons detailed in table 2, and the results from data collection, this outcome is not surprising.

Trial 3 shows that the turbidity behaved similarly to the dissolved oxygen, and that the variation introduced by it had little correlation to the FRC. From table 2, we were hoping to see that using the turbidity in conjunction with the ORP may have added value, but within PCA this did not happen. We can see that in both the DO and Tb trials, the ORP was given smaller weights than when they weren't included, so it is not surprising that our correlations were not as good.

ORP Range	**Parameters**	**Weightings of best PC**	**|COR|**	**Classification**
$< 500mV$	Cd, pH, ORP	[0.42 -0.23 0.88]	0.452	0.714
$\geq 500mV$	Cd, pH, ORP	[0.44 0.85 0.28]	0.517	0.627

Table 8: PCA Correlation (COR) and Classification Scores for the two 'linear' regions of the ORP vs FRC relationship.

Since PCA can only involve linear mappings, and we knew from figure 10 that the relationship between ORP and FRC was definitely not linear, but could be seen as two 'linear' regions, we looked at the effectiveness of using PCA in these two regions. In this case, the classification was based on whether the FRC was predicted to be too low for the low ORP group ($< 0.2mg/L$), or if it were too high for the high ORP group ($> 2mg/L$). Unfortunately, the correlation scores ended up not being very high. For the low ORP group, we still have a high weight on the ORP, which is expected due to the strong, steep linear trend for this region in the ORP vs FRC relationship. The lack of a strong correlation is because the data had too much variance unrelated to the FRC.

Surprisingly, the ORP was given a low weight in the high ORP group. This could be due to the very shallow trend that we have in this linear relationship. The pH was instead very highly weighted, which is unexpected given that we did not see any strong relationship between the pH and FRC directly. We did see some correlation between the conductivity and FRC in the data collection, but the conductivity was still not given a large weight. However, the correlation score is also not very high, so this is not providing strong support for the weightings used in this method. High conductivity and ORP weightings may have performed better, but PCA did not find enough variance along those weights to identify it.

5.4 Conclusions

Our PCA performed best when including the conductivity, pH and ORP parameters. From our data collection, we expected these parameters to have the biggest impact on our analysis. We saw that including parameters such as dissolved oxygen and turbidity had a negative effect on our results, which we associate with variances not connected with FRC. Unfortunately, the scores obtained did not suggest that this method would work well enough to give a strong prediction of the FRC. This is likely due to variance in water samples containing too much variance that isn't associated with the changing FRC, so the analysis does not align well with the objective. With more data and some exploration of other classification methods that are possible to use, there could be improvement made on the current scores given, but the evidence shown in this report suggests that PCA is not the best tool for this problem.

6 Linear Discriminant Analysis (LDA)

6.1 Background

LDA is similar to PCA in that it involves creating matrices from the data, which can be decomposed into eigenvectors that can provide dimensionality reduction, but the matrices made and the objectives of the methods are different. LDA is a supervised learning method, so each datapoint in the training set is labelled with its class. Instead of trying to maximise the variation of datapoints along an axis, it aims to maximise the separability of different classes in the dataset. The eigenvectors align to maximise variation between classes. It maximises the separation between classes in two ways:

1. Maximise the difference in means of classes

2. Minimise the variation (scatter) within a class

When we have more than 2 categories, we find a central point in the data, and maximise the distance of means to that point. LDA is a Gaussian Maximum-Likelihood Classification Technique. This means each class is assumed to have a gaussian distribution of points, with a mean and variance; classification is built into the method because datapoints can be classified based on which gaussian they are most likely to come from (which gaussian gives the maximum probability for that datapoint). In addition, LDA produces eigenvectors, that can themselves be used to produce a continuous mapping from datapoints to a predicted FRC.

LDA is supervised so we can train it to specifically find a direction aligned with changing FRC, since our classes are defined by different FRC ranges. This means our resulting eigenvectors (Linear Discriminants (LDs)) might align better with the variation due to FRC itself than in PCA.

One notable issue with LDA is that it assumes normally distributed data, statistically independent features, and identical covariance matrices for every class. This is not the case for our problem, as we have seen the ORP variation is different for different classes, and we can't assume that our variables are independent. However, this only applies for LDA as a classifier and LDA for dimensionality reduction can also work reasonably well if those assumptions are violated [23].

6.2 Methods of Implementation

Similar to PCA, we started with a decomposition just involving the conductivity, pH and ORP parameters, and then looked at how the dissolved oxygen and turbidity measurements affected the output. As opposed to PCA, including more measurements in the training should not adversely affect the output, as the method only aligns with maximum variance between classes. Therefore, if the turbidity does not provide variance attributed to the FRC, LDA should ignore the contribution from it (place a small weighting on the turbidity in the eigenvector). Another important distinction between PCA and LDA is that in LDA, the best aligned eigenvector that is obtained from the decomposition will always be the first linear discriminant (LD1); it is a supervised learning technique so it is purposefully aligning with the FRC. Because of this, we only ever need to look at LD1 (and maybe LD2) to see how effective the analysis was.

6.3 Results and Discussion

Table 9 shows the resulting eigenvectors for LDA over the whole dataset using pH, conductivity and ORP in a 3-class split. Our correlation and classification scores are quite high, but not enough to make the method reliable. Figure 14 shows the difficulties in separating the classes using the LDs found, by plotting the actual classifications against their predictions. We see that the actual classifications maintain a significant overlap between classes, particularly in the (0 → 1) region of LD1, which has prevalence of all three

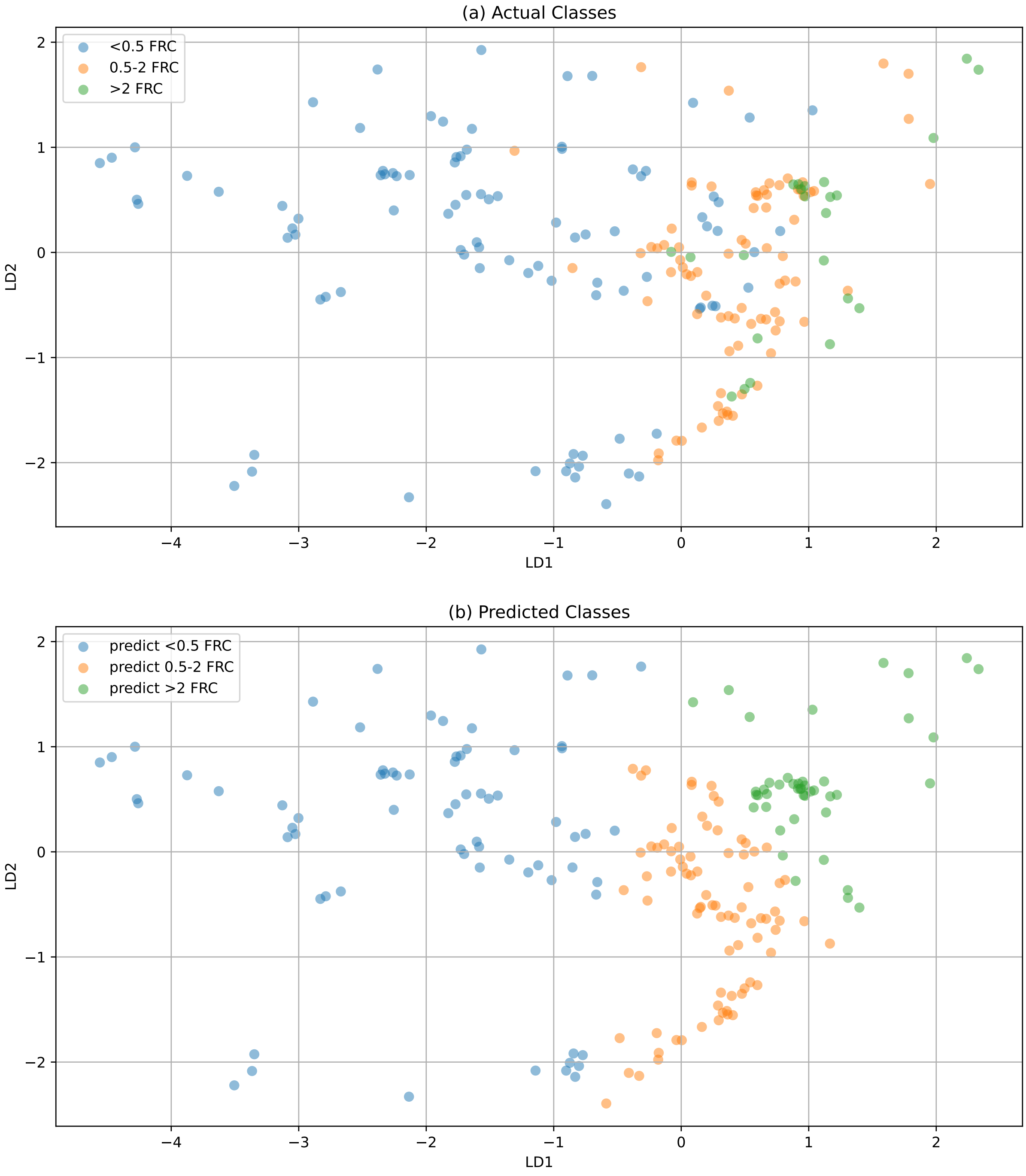


Figure 14: LDA results with data split into 3 classes using Conductivity, pH and ORP parameters. (a) shows the training labels on the datapoints plotted, and (b) shows the labels predicted by the LDA.

	Vector	COR	Classification
LD1	[0.18 0.29 0.94]	0.691	0.690
LD2	[0.77 0.61 -0.21]	-0.024	

Table 9: Initial LDA Correlations between LDs found and the FRC of All Datapoints for a 3-Class split, within a 0 to 2.5 FRC range and including DO, Tb and Cd parameters. The classification is done using both LDs, which is signified by the merged cell. These values relate to the LDs in figure 14

classes, although mainly from the (0.5 → 2 FRC) class. This overlap is what the method is trying to avoid, as it cannot distinguish between classes. However, this decomposition was the best achievable by the method, meaning the overlap was found to be very difficult to remove.

Figure 14 also shows us that there seems to be a non-insignificant difference between the direction to separate class 1 from class 2, compared to the direction to separate class 2 from class 3. We attribute this to the stark change in gradient in the ORP curve. From figure 10, it is evident that ORP is more helpful at distinguishing between $FRC < 0.5$ vs $FRC > 0.5$, compared to distinguishing $FRC < 2$ vs $FRC > 2$.

Some of the datasets hold a significant number of the overlapping points we see. One of these is the pool water samples, which contains a number of class 1 points with a value greater than 0 in LD1. It also contains a large number of the points in the cluster seen in 0.5 → 1 range for LD1 and around 0.75 in LD2. Looking at the data collection results, we see that the pool water was among the groups with high mean conductivity, and the plateau in ORP can also be seen to be quite high. These characteristics could explain why we see typically higher values (in the LD1 direction) for the class 1 points, and why the class 2 and class 3 points get clumped.

The dataset ‘Kenya Invonangya 1’ was also a big contributor, and can be explained in a similar way to the pool water, due to having a high mean conductivity and high ORP in the plateau. Since this dataset is representative of a source that might use the FRC sensor, this indicates that the method would be less reliable.

The tap water data also contributed to some of the datapoints from class 1 with higher values in LD1. This is surprising considering the low mean conductivity. However, we see in the ORP graph that there are some tap water datapoints with a high ORP but less than 0.5 FRC, which most likely contributed to these overlapping datapoints. Although we see distilled water datapoints in the same region of the ORP graph, they had an extremely low conductivity, and also lower pH than the tap water, which would be why they don’t overlap. Many of the (low FRC, high ORP) datapoints come from the tap water and distilled water samples. The distilled water is a very simple solution, and the tap water is likely to have a simpler composition than water in Kenya. This could mean that these datapoints are not so representative of values we would expect in water samples we treat. However, we would ideally still want to represent this range of water samples, so they are still relevant in evaluating the performance of our classification.

	Vector	COR	2-Class
LD1	[0.13 0.25 0.96]	0.713	0.863

Table 10: Initial LDA Correlations between LDs found and the FRC of All Datapoints for a 2-Class split, within a 0 to 2.5 FRC range and including DO, Tb and Cd parameters.

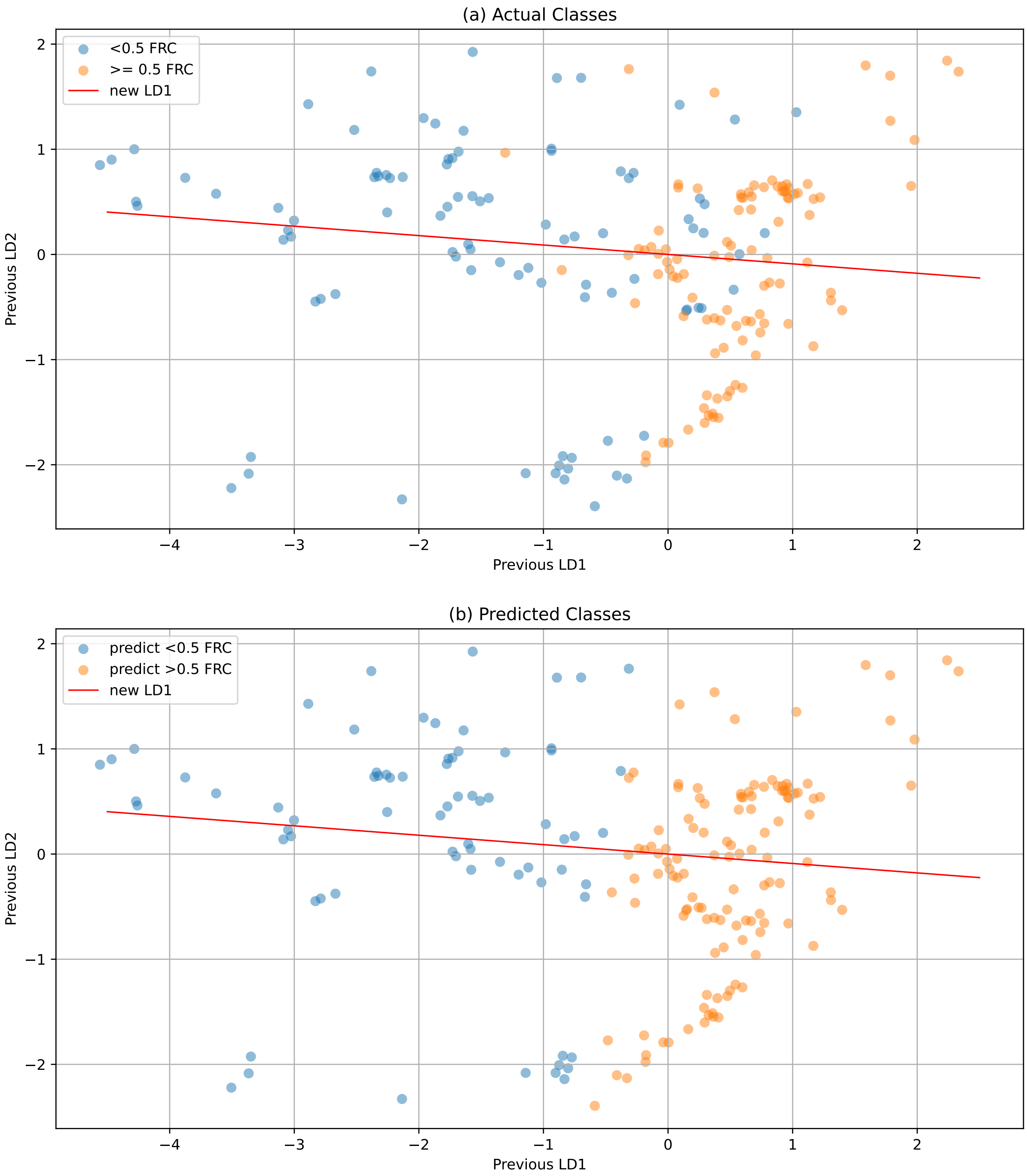


Figure 15: LDA results with data split into 2 classes, using 0 to 2.5 FRC range and Conductivity, pH and ORP. This has been plotted on the LDA directions found in the 3 class-split for easy comparison, with the LD from the 2-class split plotted. (a) shows the training labels on the datapoints plotted, and (b) shows the labels predicted by the LDA.

Table 10 shows the resulting eigenvectors for LDA over the whole dataset using pH, conductivity and ORP in a 2-class split. Interestingly, we see a slightly improved correlation of this eigenvector compared to the 3-class split. The difference between the two eigenvectors is a slight increased weight on the ORP for the 2-class eigenvector. Figure 15 shows the new LD1 plotted in the space of LD1 and LD2 found from the 3-class split. This change in direction supports what we noticed about the difference in directions found for separating different sets of classes in figure 14. We are more concerned about the distinction between class 1 and class 2, compared to class 2 and class 3, so using LD1 from the 2-class split makes sense as well.

Classification on 2-classes vs 3-classes is always going to be easier, so it is not surprising that this classification score is higher. That being said, the accuracy of the classification, although still not as high as we would hope in order to use it in the field, is still a promising indicator of the potential of this method.

Trial	**Split**	**Parameters**	**Weightings of LD1**	**|COR|**	**Classification**
1	3	Cd, pH, ORP	[0.18 0.29 0.94]	0.70258	0.659
2	2	Cd, pH, ORP	[0.13 0.25 0.96]	0.70261	0.852
3	3	Cd, pH, DO, ORP	[0.14 0.42 0.11 0.89]	0.68305	0.647
4	2	Cd, pH, DO, ORP	[0.09 0.42 0.13 0.89]	0.68306	0.843
5	3	Cd, pH, Tb, ORP	[0.17 0.28 0.04 0.94]	0.69108	0.666
6	2	Cd, pH, Tb, ORP	[0.13 0.26 0.03 0.96]	0.69112	0.862

Table 11: LDA Correlation (COR) and Classification Scores for Different Settings. Each trial is done on a 3-class or 2-class split. The 'weightings' are the average magnitude (over the 50 runs) placed on each measured parameter in LD1.

Table 11 summarises the average scores obtained for different settings of LDA conducted. Across all settings, we see that the average correlation obtained is extremely similar between the 2-class and 3-class splits, although it is always slightly higher for the 2-class split. This is most-likely due to the difference in direction highlighted from figures 14 and 15. In trials 1 and 2, we again see a very high weight placed on the ORP. Unfortunately, we also again see that the influence of including the dissolved oxygen and turbidity have been mostly negative. The one positive influence has been on the average classification score of the 2-class split including the turbidity. This is therefore providing further evidence that these measurements are not providing useful information for determining the FRC, especially considering the weights on dissolved oxygen and turbidity are all small.

7 Quadratic Discriminant Analysis (QDA)

7.1 Background

QDA is a variation of LDA, in which each class has their own covariance matrix, rather than assuming that all classes have identical covariance matrices. Because of this, QDA allows us to find curved decision boundaries rather than only linear decision boundaries, as was the case for LDA. However, the disadvantage is that QDA cannot be used for dimensionality reduction and can't be used to map to a continuous measurement like LDA or PCA, as it does not output eigenvectors aligned with the separation. Nevertheless, the curved decision boundaries mean that it has the potential to provide better classification.

7.2 Methods of Implementation

This was very similar to LDA, as the settings that can be changed to explore performance, such as number of classes or parameters included, are similar between both methods.

However, due to QDA not outputting any eigenvectors, there was no way of finding some direction that could be correlated to the FRC. As a result, the only thing to compare is the classification scores.

7.3 Results and Discussion

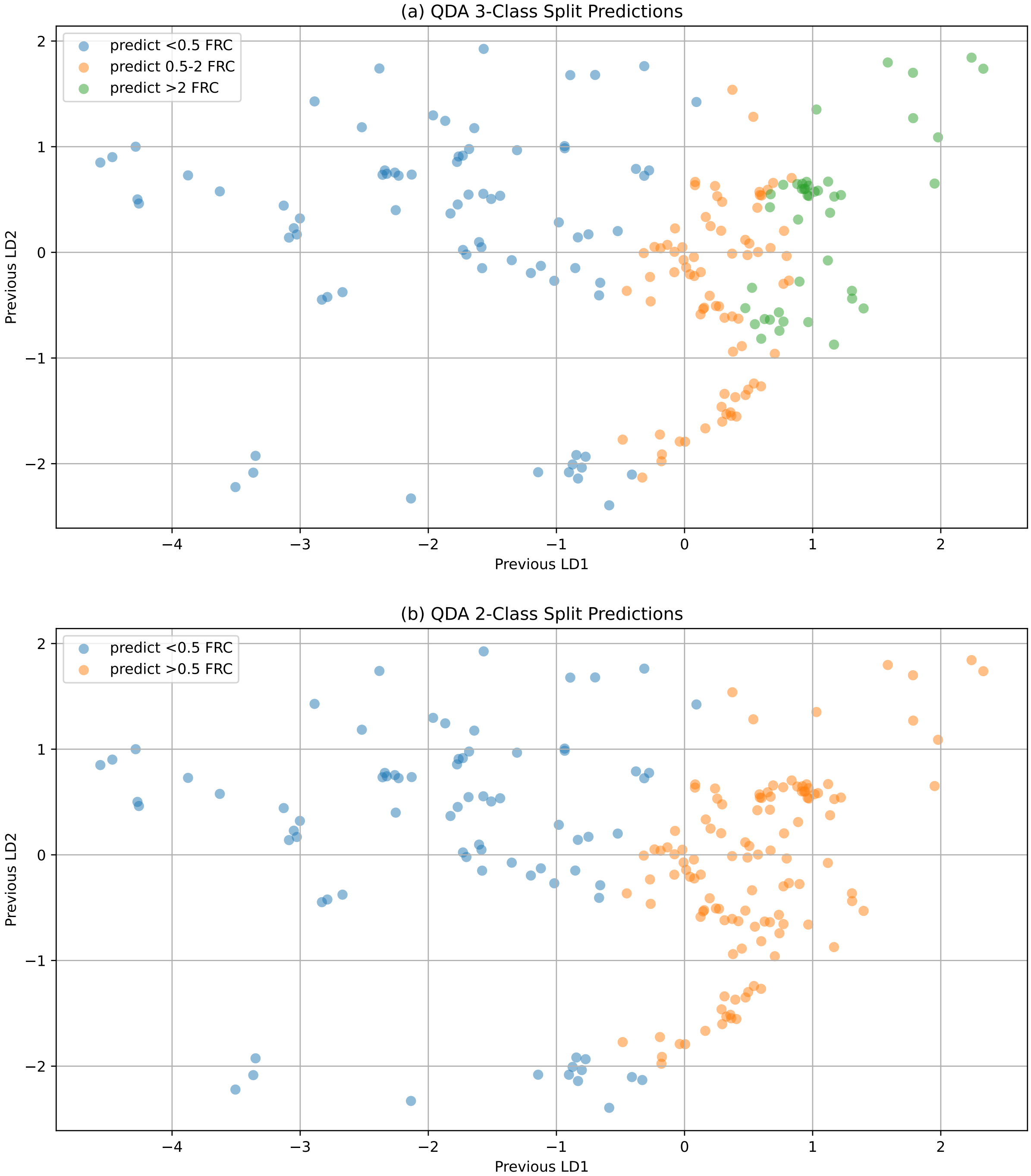


Figure 16: QDA results for 3-class and 2-class splits of the data. These have been plotted on the eigenvectors found in the 3-class LDA from figure 14, to allow for easy comparison.

Table 12 shows the resulting classification scores using QDA in a 3-class and 2-class split on the whole dataset with the pH, conductivity and ORP parameters included.

	3-Class Score	**2-Class Score**
All Data	0.777	0.893

Table 12: Initial QDA Classification Scores for a 3-Class and 2-Class Split over the Whole Dataset.

We see that both scores are quite high, with the 2-class score even being close to our goal classification accuracy of 0.9. It is important to note that these scores are still far from what we would need to justify direct implementation. Nevertheless, they are still promising indicators for the method.

Because QDA does not output any vectors related to the decomposition, we have plotted the classifications made by QDA on the LDA directions found from the 3-class split, so we can compare the decision regions. One thing we see is that QDA is able to have non-linear decision boundaries, which allows it to overcome the overlapping regions slightly better. This seen in the improved classification scores.

Trial	**Parameters**	**3-Class Score**	**2-Class Score**
1	Cd, pH, ORP	0.762	0.891
2	Cd, pH, DO, ORP	0.728	0.865
3	Cd, pH, Tb, ORP	0.74	0.873

Table 13: QDA Classification Scores for Different Included Parameters under 2-Class and 3-Class Splits.

Table 13 summarises the average classification scores for 2-class and 3-class splits on data including different sets of parameters. We see that the high classification scores have been mostly retained despite not splitting into training and test data, which is a positive indicator for using this method. We also again see reduced scores when including the dissolved oxygen and turbidity parameters. This is a strong indication that these parameters are not useful for determining FRC.

8 Comparisons

9 Conclusions

10 Future Development

- collect more data from a wider variety of water sources, preferably of closer nature to the water in question. - construct a prototype sensor to implement in the field

11 Risk Assessment

References

[1] Benjamin F Arnold and John M Colford. "TREATING WATER WITH CHLORINE AT POINT-OF-USE TO IMPROVE WATER QUALITY AND REDUCE CHILD DIARRHEA IN DEVELOPING COUNTRIES: A SYSTEMATIC REVIEW AND META-ANALYSIS". In: (2007). URL: http://www.who.int/household_water/en.

[2] Becca Bartleson. *What Is a Conductivity Meter?* Mar. 2018. URL: https://sciencing.com/conductivity-meter-5134852.html.

[3] Emerson Process Management. *Fundamentals of ORP Measurement*. May 2008. URL: https://www.emerson.com/documents/automation/application-data-sheet-fundamentals-of-orp-measurement-rosemount-en-68438.pdf.

[4] Environment and Natural Resources. *Dissolved Oxygen (DO)*. URL: http://www.h2ou.com/h2wtrqual.htm#Oxygen.

[5] *FCL Amperometric Free Chlorine Sensor — Sensorex*. URL: https://sensorex.com/product/fcl-free-chlorine-sensor/.

[6] *Handbook of water purity and quality - University of Liverpool*. URL: http://link.liverpool.ac.uk/portal/Handbook-of-water-purity-and-quality-edited-by/E49anBcvEV8/.

[7] *How to remove chemicals from water — Waterlogic*. URL: https://www.waterlogic.com/en-gb/resources/whats-in-my-tap-water/how-to-remove-chemicals-from-water/.

[8] Jiye Jin et al. "A Miniaturized FIA System for the Determination of Residual Chlorine in Environmental Water Samples". In: *Analytical Sciences* 20.1 (2004), pp. 205–207. ISSN: 0910-6340. DOI: 10.2116/ANALSCI.20.205.

[9] Sher Bahadar Khan and Kalsoom Akhtar. *Photocatalysts - Applications and Attributes*. IntechOpen, Mar. 2019, p. 156. ISBN: 978-1-78985-476-3. DOI: 10.5772/INTECHOPEN.75848. URL: https://www.intechopen.com/books/7478.

[10] M. W. LeChevallier, T. M. Evans, and R. J. Seidler. "Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water." In: *Applied and Environmental Microbiology* 42.1 (1981), p. 159. ISSN: 00992240. DOI: 10.1128/AEM.42.1.159-167.1981. URL: /pmc/articles/PMC243978/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC243978/.

[11] Peng Li. "Concise review on residual chlorine measurement: Interferences and possible solutions". In: *Journal of Cleaner Production* 323 (Nov. 2021), p. 129119. ISSN: 0959-6526. DOI: 10.1016/J.JCLEPRO.2021.129119.

[12] Peng Li, Takeshi Furuta, and Takuya Kobayashi. "Micro-particles as interfering substances in colorimetric residual chlorine measurement". In: *Ecotoxicology and Environmental Safety* 207 (Jan. 2021), p. 111279. ISSN: 10902414. DOI: 10.1016/J.ECOENV.2020.111279. URL: https://doi.org/10.1016/j.ecoenv.2020.111279.

[13] Vadim B Malkov. "Comparison of On-line Chlorine Analysis Methods and Instrumentation Built on Amperometric and Colorimetric Technologies". In: (2009).

[14] Robert F. McCracken and Mary D. Walsh. "Observations on the McLean-Van Slyke iodometric method for the titration of small amounts of halides, in its application to chlorides". In: *Journal of the American Chemical Society* 39.11 (Nov. 1917), pp. 2501–2506. ISSN: 15205126. DOI: 10.1021/JA02256A029/ASSET/JA02256A029.FP.PNG{_}V03. URL: https://pubs.acs.org/doi/abs/10.1021/ja02256a029.

[15] Amer Mohamad ' and Adnan Khayyat. "Study of Point of Use Treatment Methods for the Disinfection of Drinking Water in Nepal". In: ().

[16] National Research Council (US) Safe Drinking Water Committee. *Drinking Water and Health: Volume 2*. Washington DC: National Academies Press (US), 1980. URL: https://www.ncbi.nlm.nih.gov/books/NBK234590/.

[17] *Oxidation-reduction potential - Fact sheets*. URL: https://www.health.nsw.gov.au/environment/factsheets/Pages/orp.aspx.

[18] Nicholas G. Pizzi and American Water Works Association. *Water treatment*. 4th ed. Denver, Colorado: American Water Works Association, 2010, p. 512. ISBN: 1-61300-108-8.

[19] ProAutomation. *PoolSense — PoolSchool - Difference between ORP and Chlorine/ppm*. URL: https://www.proautomation.co/difference-between-orp-and-chlorine-ppm/.

[20] Process Instruments. "Comparison of ORP (Redox) sensors and ppm (amperometric parts per million) sensors for Swimming Pool Controls". In: (). DOI: 10.1002/9780470561331. URL: www.processinstruments.co.uk.

[21] Jukka Sassi et al. "Experiments with ultraviolet light, ultrasound and ozone technologies for onboard ballast water treatment". In: (2005). URL: http://nesteoil.com/.

[22] G. J. Schieferstein et al. "Carcinogenicity study of 3,3dimethylbenzidine dihydrochloride in BALBc mice". In: *Food and Chemical Toxicology* 27.12 (Jan. 1989), pp. 801–806. ISSN: 0278-6915. DOI: 10.1016/0278-6915(89)90111-7.

[23] Sebastian Raschka. *Linear Discriminant Analysis*. Aug. 2014. URL: https://sebastianraschka.com/Articles/2014_python_lda.html.

[24] Ian Seymour et al. "Electrochemical Detection of Free-Chlorine in Water Samples Facilitated by In-Situ pH Control Using Interdigitated Microelectrodes". In: (Feb. 2020). DOI: 10.26434/CHEMRXIV.11898126.V1. URL: https://chemrxiv.org/engage/chemrxiv/article-details/60c74857567dfef14dec49cf.

[25] Leslie Snowden-Swan, John Piatt, and Ann Lesperance. "Disinfection Technologies for Potable Water and Wastewater Treatment: Alternatives to Chlorine Gas". In: (1998).

[26] United Nations. *Scarcity — UN-Water*. URL: https://www.unwater.org/water-facts/scarcity/.

[27] US Geological Survey. *Turbidity and Water — U.S. Geological Survey*. 2018. URL: https://www.usgs.gov/special-topics/water-science-school/science/turbidity-and-water.

[28] Robert Euan Wilson, Ivan Stoianov, and Danny O'Hare. "Continuous chlorine detection in drinking water and a review of new detection methods". In: *Johnson Matthey Technology Review* 63.2 (Apr. 2019), pp. 103–118. ISSN: 20565135. DOI: 10.1595/205651318X15367593796080.

[29] World Health Organization. *Rapid Assessment of Drinking Water Quality. Country Report Nigeria*. Ed. by World Health Organization. 2012. ISBN: 789241504683. URL: https://www.who.int/publications/i/item/789241504683.