Bacterial populations and human pathogens in irrigation water sources

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

EMILY JAYNE WHITE

B.Env.Sys (Hons I), The University of Sydney

School of Life and Environmental Sciences

Faculty of Science

The University of Sydney, Australia

December 2022

Statement of Originality

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Emily J. White

Acknowledgement of Country

I acknowledge the Darug people as the Traditional Owners and Custodians of Yandhai and Dyarubbin, the Nepean and Hawkesbury Rivers.

I recognise their enduring connection and commitment to land, water, and culture, and pay my respects to Elders past and present.

Acknowledgements

This research was conducted within the Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for Food Safety in the Fresh Produce Industry (grant number: IC160100025) funded by the ARC, industry partners from Australia and New Zealand, and the University of Sydney.

The biggest thank you to my supervisors Floris van Ogtrop and Hannah Sassi. Thank you for your endless patience over the past five and a half years. Your guidance, support, and commitment to my project went above and beyond and I am truly grateful for all that you have taught me. Thank you also to my auxiliary supervisors Channah Rock and Robyn McConchie, who helped guide the direction of this project at its inception.

Thank you to WaterNSW for providing water quality data for analysis in Chapter 2. Thank you also to my colleagues at WaterNSW for encouraging me to finish writing my thesis in the last twelve months, especially my manager Lisa Hamilton for being supportive and flexible with my work arrangements, and Andrew Ball for being a wealth of knowledge about microbial risk.

Thank you to Sydney Water for hosting my internship and for resourcing Chapter 5. Thank you also to my former colleagues in the laboratories. I owe a great deal of thanks to the Microbiology team for the genuine care with which you all helped me grow as a microbiologist. I would especially like to thank Peter Beatson and Jen Ison for giving me the opportunity to join the team as an intern and then as a staff member, and Mel Wilding for her guidance and friendship.

Thank you to the University of Sydney Informatics Hub for their work on compiling the data in Chapter 2, the Ramaciotti Centre for Genomics at the University of New South Wales for the next-generation sequencing in Chapters 3 and 4, and to David Gordon at the Australian National University for generously providing the *E. coli* bloom isolates and valuable advice for Chapter 5. Thank you to the land managers who provided access to the sampling sites, and the lab managers and technicians who ensured I had access to and training for the equipment I needed. Thank you to Rob Stirzaker who accompanied me on many field trips, and James Moloney and Alex Buzacott for ensuring I was always on time for lunch at ATP.

And finally, thank you to my family, friends, and partner. Undertaking this PhD has been just one of the enormous personal challenges I have faced in the last few years, and you have been by my side throughout all of it. Without your unwavering support, I would not have made it to the end of this chapter.

Sincerely, thank you all.

Abstract

Poor microbial water quality has been implicated as the cause of foodborne disease following the consumption of contaminated fresh produce. Despite the established link between irrigation water and human health in scientific literature, current regulations and certification standards in Australia do not yet have a consistent and specific evidence-based approach to defining “safe” water sources. Due to the dynamic nature of water quality, microbial organisms, and the food production environment, the fresh produce industry also faces a considerable challenge in monitoring water for human health risks within existing assurance frameworks.

The United States fresh produce industry has developed a tool for predicting microbial contamination within irrigation water sources. The University of Arizona’s ‘AgWater App’ determines the likelihood of coliform or *Escherichia coli* bacteria in water under current environmental conditions and using historical data, and therefore the likelihood of microbial contamination within a water source exceeding the Food Safety Modernization Act (FSMA) standards. Freshcare is Australia’s most common fresh produce safety standard and is also the only local assurance program with prescriptive water quality criteria. Under certain high-risk conditions, Freshcare requires preharvest water to meet *E. coli* <100 cfu per 100mL. Using the ‘AgWater App’ approach and the Freshcare criterion, this thesis determines significant correlations between environmental and physicochemical water quality and *E. coli,* and consequently evaluates the validity of statistical models predicting irrigation water risk. The results of this analysis stress the importance of considering sediment-water exchanges in agricultural water sources, and reiterate the impact of rainfall, temperature, and turbidity on water quality. Also highlighted by this research is the difficulty of overcoming site-specific influences when trying to design a tool that can be applied widely.

Arising from this topic of study is the inevitable examination of *E. coli* as a suitable indicator of microbial water quality. To answer the research question “Does the faecal indicator bacterium *E. coli* capture all microbial risks associated with fresh produce irrigation water?”, this thesis generated metagenomic 16S rRNA data for consortia of bacteria that were likely to be largely unculturable. Taxonomic and functional profiles of water and sediment communities were explored to better understand the range of human health-related taxa present in irrigation water sources. It was revealed that a considerable number of bacteria related to food safety concerns were present in the environmental samples, including Bacteroides, Cyanobacteria, and Proteobacteria. Further, inferred functional pathways implicated in infectious diseases from *Legionella*, *Salmonella*, *Staphylococcus*, *Vibrio*, and *Shigella* were identified. Additional investigation into spatial and temporal trends in the

metagenomic data showed that there was seasonal variation and sediment-water exchanges that support similar findings in other studies. Exploring the metagenomic data further, this thesis used a novel measure of community cohesion to show that food safety risks in environmental water sources are ultimately dependent on bacterial community dynamics.

The utility of *E. coli* as an indicator for Australian water sources was further evaluated in this thesis by validating a diagnostic protocol for environmental, bloom-forming strains and applying it to the water and sediment samples. Though bloom strains were not detected in the study area, this work remains a useful consideration for growers who may be inadvertently disqualifying irrigation water sources based on inflated *E. coli* measurements.

Overall, this thesis contributes to the ongoing discourse on the challenges associated with assigning risk to irrigation water sources and provides food for thought on the future application of metagenomics in produce safety.

Table of Contents

1 An introduction to fresh produce irrigation water quality....................................... 1

1.1 Introduction 1

1.2 Irrigation water and human health 2

1.2.1 Sources of microbial contamination 2

1.2.2 Microbes of concern 3

1.2.2.1 Bacteria 3

1.2.2.2 Viruses 4

1.2.2.3 Protozoa 5

1.2.3 Foodborne disease outbreaks linked to irrigation water 5

1.3 Current recommendations for irrigation water use 8

1.3.1 Water sources 8

1.3.2 Irrigation application and timing 8

1.3.3 Water quality requirements for certification 9

1.4 Monitoring water quality for food safety 10

1.4.1 Indicators of water quality 10

1.4.2 Targeted molecular assays and next-generation advancements 11

1.4.3 Forecasting and modelling 12

1.5 Thesis overview 13

2 Environmental and water quality parameters as predictors of *Escherichia coli* in irrigation water sources 15

2.1 Introduction 15

2.2 Methods and materials 16

2.2.1 Hawkesbury dataset 16

2.2.1.1 Field sampling and testing 17

2.2.1.2 Laboratory analysis 18

2.2.2 Greater Sydney dataset 19

2.2.3 Statistical analysis 19

2.2.3.1 Hawkesbury dataset 20

2.2.3.2 Greater Sydney dataset 20

2.3 Results 21

2.3.1 Hawkesbury dataset 21

2.3.1.1 Climate variables ... 21
2.3.1.2 Site characteristics ... 21
2.3.1.3 Comparison of sediment and water ... 25
2.3.1.4 E. coli contamination ... 25
2.3.1.5 Generalised additive model (GAM) ... 27
2.3.2 Greater Sydney dataset ... 31
2.3.2.1 Data summary ... 31
2.3.2.2 Regression model ... 32
2.3.2.3 Classification model ... 34
2.4 Discussion ... 36
2.4.1 Sediments ... 36
2.4.2 Spatiotemporal variation and climate factors ... 37
2.4.2.1 Site-specific influences ... 38
2.4.3 Physicochemical variation ... 39
2.4.4 Model performance ... 40
2.4.5 Limitations and further research ... 40
2.5 Conclusion ... 41
3 Bacterial community composition and food safety risks in irrigation water sources ... 42
3.1 Introduction ... 42
3.2 Methods and materials ... 43
3.2.1 Sampling ... 43
3.2.2 Sample preparation and DNA extraction ... 43
3.2.3 Next-generation sequencing (NGS) and data processing ... 44
3.2.4 Taxonomic and functional profiling ... 45
3.2.5 Irrigation pathogen potential index (IPPI) ... 45
3.2.6 Data analysis ... 46
3.3 Results ... 46
3.3.1 Metagenomic sequence generation ... 46
3.3.2 Taxonomic profiles ... 48
3.3.3 Irrigation pathogen potential index (IPPI) ... 51
3.3.4 Functional profiles ... 54
3.4 Discussion ... 56

3.4.1 Diversity ... 56
3.4.2 Taxonomic profiles ... 57
3.4.3 Irrigation pathogen potential index (IPPI) ... 58
3.4.4 Functional profiles ... 58
3.4.5 Limitations ... 59
3.5 Conclusion ... 61
4 Dynamics of bacterial populations and human pathogens in irrigation water sources ... 62
4.1 Introduction ... 62
4.2 Methods ... 63
4.2.1 Sampling and previous analyses ... 63
4.2.2 Alpha diversity ... 63
4.2.3 Beta diversity and PERMANOVA ... 63
4.2.4 Connectedness and cohesion ... 64
4.2.5 *E. coli* and community characteristics ... 65
4.3 Results ... 65
4.3.1 Alpha diversity ... 65
4.3.2 Beta diversity and PERMANOVA ... 66
4.3.3 Connectedness and cohesion ... 71
4.3.4 *E. coli* and community characteristics ... 74
4.4 Discussion ... 76
4.4.1 Alpha diversity ... 76
4.4.2 Beta diversity and PERMANOVA ... 76
4.4.3 Connectedness and cohesion ... 77
4.4.4 *E. coli* and community characteristics ... 78
4.4.5 Limitations ... 78
4.5 Conclusion ... 79
5 Environmental, bloom-forming *Escherichia coli* ... 80
5.1 Introduction ... 80
5.2 Characterisation of a bloom isolate collection ... 84
5.2.1 Isolate information ... 84
5.2.2 Growth and morphology characteristics ... 85
5.2.2.1 MacConkey and tryptic soy agar (35°C) ... 85

5.2.2.2 m-FC agar (35°C; 44.5°C) 86
5.2.2.3 Tryptic soy agar (44.5°C) 88
5.2.3 Enzyme tests 88
5.2.4 Biochemical profiles 89
5.2.5 Capsule typing 90
5.2.6 Isolate characterisation summary 91
5.3 Validation of the diagnostic assay at Sydney Water 92
5.3.1 East Coast pool 92
5.3.2 West Coast pool 93
5.3.3 Optimised protocol 93
5.4 Application on pooled environmental samples 94
5.4.1 Spiked environmental samples 94
5.4.2 Screening of the upper Hawkesbury River 95
5.5 Discussion and conclusion 96
5.5.1 Future research 96
5.5.1.1 Dye-based assay 97
5.5.1.2 Probe-based assay 97
5.5.1.3 Metagenomics 97

6 Discussion 98

6.1 What was the extent of microbial contamination in the studied irrigation water sources? 98
6.1.1 How did the sediment contribute to the microbial loads? 99
6.1.2 Which environmental and physicochemical water quality parameters were related to increased contamination? 99
6.2 Are the studied water sources suitable for the irrigation of fresh produce? 99
6.3 What additional information did next-generation sequencing contribute? 100
6.4 Was *E. coli* a useful indication of microbial risk? 102
6.5 Recommendations 102
6.5.1 Microbial limits and risk assessment 102
6.5.2 Monitoring frequency and data collation 103
6.6 Conclusion 104

7 Supplementary material 105

8 References 108

List of Figures

Figure 1.1. Conceptual diagram of fresh produce contamination arising from pathogens in pre-harvest water sources 2

Figure 2.1: Field sampling sites in the upper Hawkesbury River area, New South Wales. River sites (n = 5) are marked blue; on-farm water storages (n = 3) are marked red. The Hawkesbury River begins at the confluence of the Grose and Nepean Rivers (left) and flows north-east towards the Pacific Ocean. 17

Figure 2.2: WaterNSW sampling sites in the Hawkesbury-Nepean River system, Greater Sydney, NSW. The Nepean River flows north and becomes the Hawkesbury River at its confluence with the Grose River in north-west Sydney. 19

Figure 2.3: Definitions and equations of the statistics used to evaluate the performance of the random forest classification model (i.e., *E. coli* concentrations exceeding the Freshcare irrigation water requirement). 21

Figure 2.4: A) daily rainfall (mm), B) daily maximum temperature (°C), and C) daily global solar exposure (MJ m^{-2}) at BoM station #67105 (Richmond RAAF) during the period of field sampling, i.e., spring 2018 to winter 2019 22

Figure 2.5: A scatterplot of log-transformed *E. coli* concentrations in the sediment (x-axis) and the water (y-axis) with the fitted Akritas-Theil-Sen (ATS) trendline in purple (intercept 3.58; slope 0.23) 25

Figure 2.6: Boxplots of *E. coli* concentrations (MPN per 100mL) in the water samples (n = 96) by site. 26

Figure 2.7: Boxplots of *E. coli* concentrations (MPN per 1g) in the sediment samples (n = 94) by season. 27

Figure 2.8: The response of log10 *E. coli* concentration (z-axis) to daily rainfall (DF 0.99) (x-axis) and maximum temperature (DF 0.5) (y-axis), as modelled by the GAM. 28

Figure 2.9: A smooth term plot for longitude (edf = 5.86), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis. 29

Figure 2.10: A smooth term plot for electrical conductivity (edf = 3.68), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis. 29

Figure 2.11: A smooth term plot for dissolved oxygen (edf = 2.14), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis. 30

Figure 2.12: A smooth term plot for turbidity (edf = 3.47), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis. 30

Figure 2.13: Variable importance plots for the regression random forest model predicting *E. coli* concentrations. The left plot shows the percent increase in the mean square error (MSE) when the term was removed from the model. The right plot shows the increase in node purity when the term was included in the model. 32

Figure 2.14: Partial (marginal) dependence plots of the regression random forest model for latitude, rainfall (DF 0.75), and turbidity. The y-axis is the modelled response of log10 *E. coli* MPN per 100mL. The observed values used in the model calibration are indicated by tick marks along the y-axis. 33

Figure 2.15: Observed and predicted *E. coli* organisms $100mL^{-1}$ for the validation of the random forest regression model. The red line indicates a 1:1 gradient, which would be a perfect prediction of the observed values. 33

Figure 2.16: Variable importance plots for the classification random forest model of *E. coli* concentrations exceeding the Freshcare certification requirement of 100 cfu $100mL^{-1}$. 34

Figure 2.17: Partial dependence plots of the classification (binary E. coli) random forest model 35

Figure 2.18: The receiver operator characteristic (ROC) graph for the random forest model on the binary E. coli variable (i.e., exceeding the Freshcare requirement). The red line indicates the actual performance of the model, with an area under the curve of 0.86. The grey dashed line indicates a 1:1 gradient of the true positive rate (TPR) and the false positive rate (FPR), or a completely random classification; perfect classification models would have a TPR of 1 and a FPR of 0, and the ROC would end in the top left of the graph. 35

Figure 3.1: Diagram of sample preparation for DNA extraction and 16S sequencing. DNA was extracted from the sediment samples directly. The water samples underwent concentration (i.e., flocculation, centrifugation, and filtration) prior to DNA extraction. 44

Figure 3.2: Relative abundance of phyla (>5%) in the sediment and water samples (n = 190) across the four seasons from September 2018 to August 2019 season (Spring: September-November; summer: December-February; autumn: March-May; winter: June-August) 48

Figure 3.3. Relative abundance of phyla (>5%) in the sediment (n = 94) and water (n = 96) samples from on-farm storages (n = 3) and river sites (n = 5). 49

Figure 3.4: Relative abundance (%) of the top twelve orders within the Proteobacteria phylum in the water and sediment samples across the four seasons from spring, 2018, to winter, 2019. 50

Figure 3.5: Relative abundance (%) of genera within the Enterobacterales order across the four seasons in the sediment and water samples from on-farm storages and river sites 51

Figure 3.6: Boxplots of the log10-transformed irrigation potential pathogen index, separated by sample source and matrix, with significant comparisons shown. 53

Figure 3.7: Boxplots of the log10-transformed potential pathogen proportions, separated by sampling season, with significant comparisons shown. 53

Figure 3.8: KEGG pathways predicted by Piphillin, separated by sample source and matrix. 55

Figure 3.9: KEGG pathways predicted by Piphillin in the 'Infectious disease: bacterial' tier, separated by sample source and matrix. 56

Figure 4.1: Shannon (alpha) diversity by sample matrix and source type. Asterisks are indicative of pairwise significance. 65

Figure 4.2: Shannon (alpha) diversity by season (Spring: September-November; Summer: December-February; Autumn: March-May; Winter: June-August). Asterisks are indicative of pairwise significance. 66

Figure 4.3: Nonmetric multidimensional scaling (stress = 0.16) of the water and sediment samples (n = 190). The ellipses are indicative of group distributions only. PERMANOVA on the Bray-Curtis distance for the sample types was significant (P < 0.01). Analysis of the homogeneity of the dispersion for the sample types was not significant (F = 1.22; P = 0.311). 67

Figure 4.4: Nonmetric multidimensional scaling (stress = 0.048) of the water samples (n = 96). The arrows are indicative of a linear trend with the ordination scores only. 68

Figure 4.5: Nonmetric multidimensional scaling (stress = 0.041) of the sediment samples (n = 94). The arrows are indicative of a linear trend with the ordination scores only. 70

Figure 4.6: Positive cohesion values by sample matrix and source type. Asterisks are indicative of significant differences in the positive cohesion values as calculated by a Wilcoxon test. 73

Figure 4.7: Negative cohesion values by sample matrix and source type. Asterisks are indicative of significant differences in the positive cohesion values as calculated by a Wilcoxon test. 74

Figure 5.1: Gel electrophoresis of the capsule gene PCR products, individually and pooled. *galF* (668bp) indicates a Group 1 capsule is present..... 84

Figure 5.2: Example bloom isolates incubated on A) MacConkey and B) tryptic soy agar... 86

Figure 5.3: Observed variation in *Escherichia coli* isolate morphologies as grown on m-FC agar, A and B are East coast isolates; C is a West Australian isolate, and D is a laboratory reference strain. 87

Figure 5.4: Growth of bloom isolates on tryptic soy agar (TSA) after incubation overnight at 44.5°C..... 88

Figure 5.5: Electrophoresis gel showing amplicon size of the East Coast pool PCR products 92

Figure 5.6: Electrophoresis gel showing amplicon size of the West Coast pool PCR products 93

Figure 5.7: Electrophoresis gel of the spiked samples A and B, and the East Coast and West Coast pool PCR products..... 95

Figure 7.1: A map of the upper Hawkesbury River region, with irrigated land for seasonal or perennial horticulture or other crops indicated (“NSW Landuse 2017 v1.2,” 2020)..... 105

Figure 7.2: Soil texture triangle to determine textures given the sand, silt, and clay percentages of a sample..... 106

Figure 7.3: Rarefaction curves for the water and sediment samples (n = 190). Appropriate sampling depth is achieved when the rarefaction curve flattens, i.e., no further unique ASVs are generated from sampling more sequences. 106

List of Tables

Table 1.1: Examples of fresh produce outbreaks linked to irrigation water 7

Table 1.2: Microbial limits for irrigation water from guidelines, legislation, and certification schemes 10

Table 2.1: Hawkesbury region field site sampling locations and important site characteristics 23

Table 2.2: Mean water quality conditions at the Hawkesbury region field sites during the period of sampling (standard deviation in brackets). DO = dissolved oxygen; PC = phycocyanin. *The values for the *E. coli* concentrations are Kaplan-Meier estimates of the restricted mean and standard deviation due to the censoring of the data 24

Table 2.3: Pairwise comparisons (Peto-Peto test) of *E. coli* concentrations in the water (* indicates significant differences, $P < 0.05$) 26

Table 2.4: Coefficients for the generalised additive model (GAM) of the concentration of *E. coli* in the water column (n = 86). The GAM explained 62.9% of deviance and had an adjusted R^2 value of 0.492. 28

Table 2.5: Water quality and climate summary statistics (median, mean, standard deviation) for the Greater Sydney dataset 31

Table 3.1: The distribution of the number of samples, amplicon sequence variants (ASVs), and sequence reads from samples (n = 190) collected between September 2018 and August 2019... 47

Table 3.2: Number of amplicon sequence variants (ASVs) successfully mapped to taxonomic ranks in the SILVA (v138) database, including percent of total ASVs................................. 47

Table 3.3: Number of 16S rRNA sequence reads, amplicon sequence variants (ASVs), relative abundance (%) of total sequence reads, and number of member genera identified in taxonomic groups across all samples (n = 190) with the SILVA (version 138) database. 47

Table 3.4: Mean relative abundances (%) of sequence reads attributed to each of the potentially pathogenic genera, separated by sample source and matrix (ND = not detected). 52

Table 3.5: Mean relative abundances (%) of sequence reads attributed to each of the potentially pathogenic genera, separated by season (ND = not detected)........................... 52

Table 4.1: PERMANOVA output for the Bray-Curtis distances of the water samples (n = 96) showing the marginal effect of terms with NA values excluded. Asterisks are indicative of significance 69

Table 4.2: PERMANOVA output for the Bray-Curtis distances of the sediment samples (n = 94) showing the marginal effects with NA values excluded. Asterisks are indicative of significance. 71

Table 4.3: The taxonomies of the ten ASVs with the highest positive connectedness 72

Table 4.4: The taxonomies of the ten ASVs with the highest negative connectedness 72

Table 4.5: Pairwise Spearman correlations for the water samples (n = 96). 75

Table 4.6: Pairwise Spearman correlations for the sediment samples (n = 94). 75

Table 5.1: Historical environmental Escherichia coli blooms in Australia, by location and year or approximate date (Bertone et al., 2019; Mackay and Ridley, 1983; Phelan, 2019; Power et al., 2005; Sinclair, 2019). 81

Table 5.2: Bloom-forming *Escherichia coli* isolate information, including identification, isolation location, phylogroup, and Group 1 capsule type 85

Table 5.3: Detailed summary of *Escherichia coli* bloom isolates, including identification, phylotype, Group 1 capsule type, primer set, expected amplicon size, colony growth at 44.5°C, MUG production, and API 20E biochemical profiles. 91

Table 5.4: Details of the primer sets used in the diagnostic assay, including target genes, and PCR product sizes 94

Table 7.1: Particle size analysis data for the sediments of the Hawkesbury field sampling sites, calculated according to the soils texture grid. All hydrometer readings corrected with a blank solution. 4:48 minute reading based off temperature of 20°C........ 105

Table 7.2: Successful species assignments for the genera included in the irrigation pathogen potential index (IPPI) 107

1 An introduction to fresh produce irrigation water quality

1.1 Introduction

Water microbiology is an important consideration for public health because people can acquire disease by consuming or contacting poor quality water. This link between water and human health underpins the operation of water utilities across the world. In general terms, the greatest microbial risk from water is the direct ingestion of water contaminated with animal or human faeces (WHO, 2019). The consequence is that the predominant global water quality problem is waterborne diseases, such as cholera, concentrated in regions where the population is underserved by water, sanitation, and hygiene infrastructure (WHO, 2019).

However, health-related water microbiology remains a concern beyond drinking water supplies in that it also relates to food safety. The overlap of water quality and food safety is the microbial contamination of irrigation water – waterborne pathogens can induce the development of foodborne diseases through the practice of irrigating fresh produce (Steele and Odumeru, 2004). In fact, an increasing number of gastrointestinal disease outbreaks related to fresh produce consumption have been traced back to the pre-harvest application of poor quality irrigation water (Bennett et al., 2020; Callejón et al., 2015; Gelting et al., 2011; Jenkins et al., 2015).

Fresh produce production is one of Australia's largest industries. In 2021, fruit and vegetable production reached a value of $10.7 billion ("Australian Horticulture Statistics Handbook 2020/21," 2022) from an area of approximately 300,000 hectares (ABS, 2022). The Australian fresh produce industry consumed approximately 1.5 million megalitres of fresh water to achieve this level of output (ABS, 2022). While epidemiological data is not available for this period, the Communicable Diseases Intelligence report for the year 2017 (OzFoodNet Working Group, 2022) described a single foodborne disease outbreak where fresh produce was the probable vehicle of infection, which resulted in 11 people reporting as ill following consumption. While this highlights Australia's excellent fresh produce food safety reputation, gastrointestinal illness is often substantially underreported (Gibney et al., 2014), and only 61% of traceback investigations were successful in that reporting year (OzFoodNet Working Group, 2022). The industry is committed to continuing to provide safe fresh produce to its consumers, and so substantial investment has been made to progress research into multiple areas of concern, including pre-harvest water quality.

1.2 Irrigation water and human health

Irrigation is an essential pre-harvest step in the production of fresh produce, and one of the first points in the production chain where microbial contamination can be introduced and therefore mitigated. As conceptually described in Figure 1.1, pathogens are excreted into the environment from infected humans and animals, where they can accumulate in water sources. If contaminated water is used for irrigation, the fresh produce may also become contaminated. Increased stress is then placed on pre- and post-harvest pathogen reduction steps to effectively remove the microbial contamination. Insufficient disinfection can lead to disease in humans when the produce is consumed. From the perspective of a grower, this cycle can be disrupted if the use of contaminated water is avoided. Water that is in contact with food should be of acceptable microbial quality.

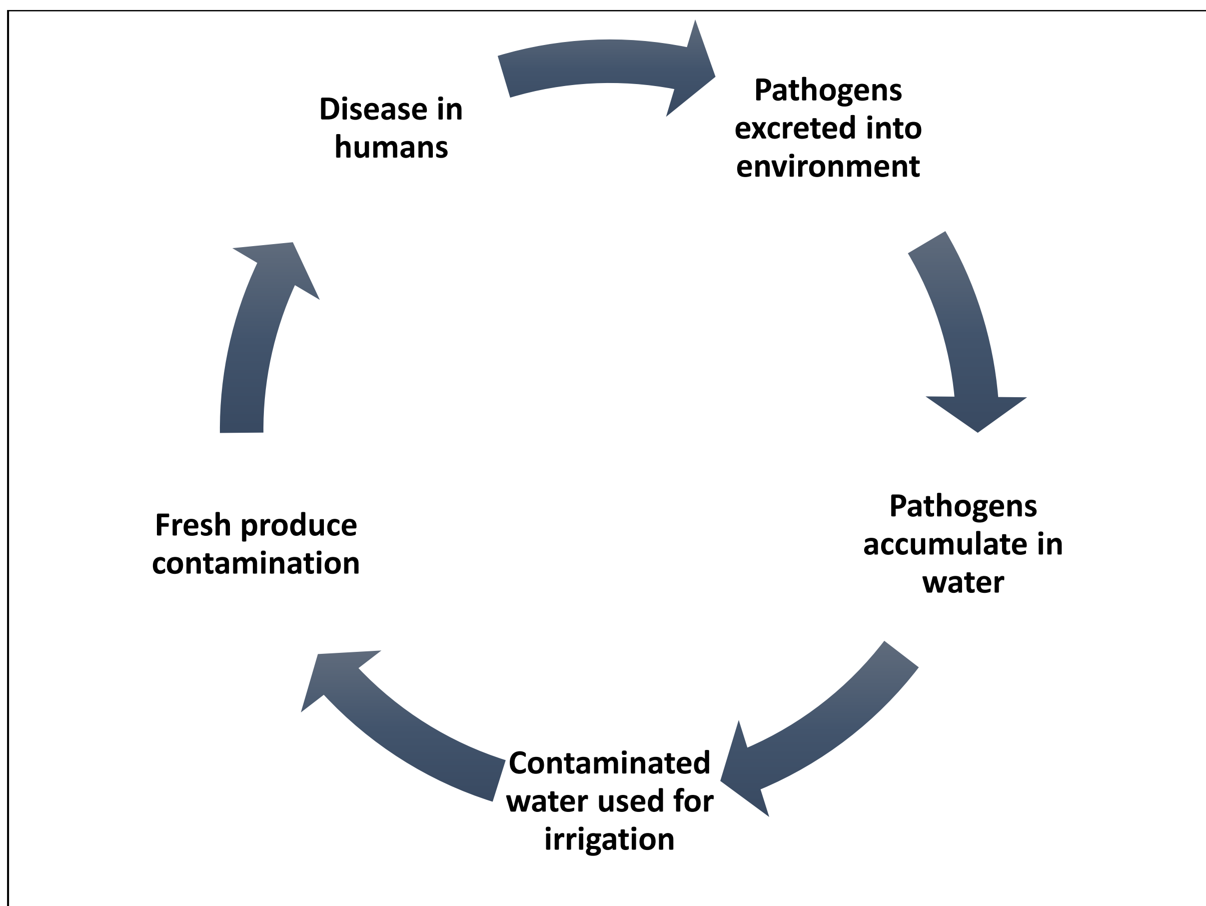


Figure 1.1. Conceptual diagram of fresh produce contamination arising from pathogens in pre-harvest water sources

1.2.1 Sources of microbial contamination

A direct pathway for water sources to become contaminated with pathogens is through exposure to faecal material from humans or animals. The extent to which these sources are contaminated usually depends on the amount of faecal material input relative to the size of

the water source, and the protective measures put in place to mitigate this contamination (NHMRC and NRMMC, 2011). Surface waters are therefore inherently more exposed to microbial contamination than groundwater sources.

Human faecal contamination in environmental water bodies can arise from several circumstances. Compromised sewage infrastructure can lead to the direct ingress of pathogens from infected humans; insufficient on-site sewage management systems (Carroll et al., 2004; Withers et al., 2014), sewage treatment plant overflows and failures (Devane et al., 2014; Sojobi and Zayed, 2022), and intentional effluent discharges (Xie et al., 2022) are several known avenues. Primary contact recreation in water bodies, i.e., swimming, can also increase the vulnerability of a water supply to faecal contamination (Gerba, 2000) and requires consideration when managing drinking water sources (NHMRC and NRMMC, 2011).

Livestock and wild animals can carry microbes capable of causing disease in humans, known as zoonotic pathogens. Effluent and run-off from intensive animal production, e.g., feedlots, dairies, poultry farms, piggeries, is a proven source of pathogenic bacteria in watersheds (Petit et al., 2017) and can increase the contamination of irrigation water sources significantly (Harris et al., 2018). The prevalence of pathogens in water sources is often exacerbated by extreme rainfall (Cann et al., 2013; Curriero et al., 2001). Direct animal and bird intrusion in surface waters is another primary cause of faecal contamination (Graczyk et al., 2008; Smolders et al., 2015; Wither et al., 2005), as is agricultural run-off from paddocks or fields treated with animal manure (Kumar et al., 2013; Wilcock et al., 2011).

1.2.2 Microbes of concern

Some key pathogenic microbes that have been detected in irrigation water or on fresh produce are summarised below. They include bacteria, viruses, and protozoa. Studies examining fresh produce safety most often focus on enteric pathogens, as these gut organisms can transfer into irrigated produce and digestion is the main route of exposure to harm.

1.2.2.1 Bacteria

Escherichia coli (*E. coli*), or *Shigella spp.* (Chattaway et al., 2017), primarily inhabit the lower gastrointestinal tract of humans and other warm-blooded animals (Luo et al., 2011; Tenaillon et al., 2010). Shiga toxin-producing *Escherichia coli* (STEC) is a pathotype of *E. coli* that can cause severe illness in humans (Paton and Paton, 1998). *E. coli* O157:H7 is a prolific foodborne outbreak agent (Machado-Moreira et al., 2019) and has been identified as the

cause of several outbreaks linked to contaminated irrigation water (Crawford and Baloch, 2010; Gelting et al., 2011; Hajmeer et al., 2007).

Salmonella, a genus from the family Enterobacteriaceae, is one of the most common pathogens responsible for foodborne disease outbreaks (Liu et al., 2018). According to the CDC, *Salmonella* outbreaks made up approximately 64% of all multistate foodborne outbreaks investigated in the USA from 2017 to 2020, with fruit and seeded vegetable consumption accounting for a large proportion of these outbreaks (CDC, 2022). Published traceback investigations have suggested that irrigation water may be a source of *Salmonella* contamination of produce (Cito et al., 2014; Greene et al., 2008; Mody et al., 2011). *Salmonella* has been found in surface waters such as rivers, lakes, and ponds (Antaki et al., 2016; Benjamin et al., 2013; Li et al., 2014), which can harbour diverse serotypes (McEgan et al., 2014).

In comparison to *E. coli* and *Salmonella*, fewer studies have been conducted on the prevalence and origins of *Listeria monocytogenes* contamination (Gartley et al., 2022). *L. monocytogenes* is thought to survive in biofilms (Rodríguez-López et al., 2018), which play an important role in contamination in packing houses and processing facilities. Several studies have shown that irrigation water sources can contain *L. monocytogenes* (Gartley et al., 2022). *Listeria spp.* can often be isolated from soil, water, and animal sources (Lyautey et al., 2007; Stea et al., 2015). As a result, *L. monocytogenes* can be found in farm environments and is an irrigation water concern.

There are a range of other bacteria that can induce gastrointestinal infection in humans when ingested. For example, some species and strains of *Aeromonas, Campylobacter,* and *Clostridium* (Gracey et al., 1982; Labbé and Juneja, 2013; Palmer et al., 1983) can cause infections that range from mild to severe and can potentially be fatal. These bacterial infections can cause a variety of symptoms, such as diarrhea, vomiting, fever, and sepsis.

1.2.2.2 Viruses

Enteric viruses such as adenoviruses, enteroviruses, norovirus, hepatitis A, and rotavirus are responsible for a substantial number of fresh produce outbreaks and individual illness each year (Machado-Moreira et al., 2019) and are known to be isolated from surface water used for irrigation (López-Gálvez et al., 2016). Viruses, due to their small size, can also move in the subsurface environment far more easily than larger microbes such as bacteria and protozoa. The ready transport of viruses through soil can lead to elevated viral presence in groundwater; a study of 23 different groundwater wells in South Korea detected human viruses (i.e., norovirus, enteroviruses, and adenovirus) in a third of the sites and bacterial concentrations over the national standard for three quarters of the sites (Jung et al., 2011). A

study of three farms in South Korea detected enteric viruses in 17% of groundwater samples and 10% of raw vegetable samples (Cheong et al., 2009).

1.2.2.3 Protozoa

The protozoa (parasites) commonly assessed in drinking water sources are *Cryptosporidium*, *Giardia*, and *Cyclospora* (NHMRC and NRMMC, 2011). These protozoa are often isolated from raw fruits and vegetables (Li et al., 2020), and are also responsible for a notable number of reported outbreaks and illnesses in fresh produce (Herman et al., 2015; Machado-Moreira et al., 2019). Protozoa are resistant to several forms of disinfection (Thomas, 2012), which is an important consideration if post-harvest sanitisation processes are to be relied upon for their effectiveness.

1.2.3 Foodborne disease outbreaks linked to irrigation water

There has been a recent, global increase in the number of foodborne disease outbreaks related to fresh fruit and vegetables (Berger et al., 2010; Olaimat and Holley, 2012). In a number of these cases, investigators identified irrigation water as the likely source of fresh produce contamination (Steele and Odumeru, 2004; Uyttendaele et al., 2015). In linking these outbreaks to irrigation water, researchers must consider a variety of environmental influences and irrigation practices in order to pinpoint a definite cause of fresh produce contamination (Gelting and Baloch, 2013). It is difficult to implicate irrigation water as the cause of fresh produce contamination, or to link it to a specific disease outbreak, because of the dynamic nature of water: by the time water contamination is suspected or apparent, the water conditions may have changed, and testing the water becomes ineffectual. Though scientific advances, such as pulsed-field gel electrophoresis (PFGE) and whole genome sequencing (WGS), have allowed investigators to confirm genetic links between environmental DNA samples and clinical patient strains (Hoelzer et al., 2018), successful outbreak investigations may still require both extensive and timely sampling of the irrigation water system (Gelting et al., 2015). The dynamic nature of water can cause water quality to change significantly in short periods of time and thus, not every outbreak is resolved (CDC, 2016). A food safety outbreak will often outpace the efforts of those trying to identify and mitigate the cause.

In 2006, an *E. coli* O157:H7 outbreak associated with fresh bagged spinach from a Californian farm affected consumers across 26 states of the US (CDC, 2006). The FDA reports that 199 people became sick, 102 were hospitalised, and three died as a result of their infection. A comprehensive investigation explored possible irrigation water-related influences at a watershed-scale (Gelting et al., 2011). There was a genetic match between the outbreak strain and water and soil samples taken from the farm environment, as

confirmed by the California Food Emergency Response Team (CalFERT) (Hajmeer et al., 2007). Due to the genetic link between the outbreak strain and the environmental samples, the CalFERT investigation hypothesised that the contaminant was likely to have transferred onto the lettuce and soil via the application of blended irrigation water. A catchment-wide study exploring the circumstances behind the produce contamination reached the same conclusion (Gelting and Baloch, 2013). The farm's irrigation water was from three sources: pumped groundwater, surface water delivered through pipes from the local water management agency, and effluent from wastewater lagoons on two neighbouring dairy farms (Hajmeer et al., 2007). The grower intended for the wastewater effluent to be used only on animal feed crops, but there was a potential for the system to fail and allow backflow through the complex piping network that was not designed for such use. The second investigation (Gelting and Baloch, 2013) also looked into the possibility that nearby dairy effluent lagoons contaminated the groundwater via hydrologic forces, which are a known pathway of contaminant transfer (Bradford and Harvey, 2016).

An *E. coli* O145 outbreak in 2010 infected 33 people (26 confirmed; 7 probable) across five US states. The cause was contaminated romaine lettuce from a farm in Arizona USA, as confirmed by the FDA (CDC, 2010). The lettuce fields were irrigated with surface water from an open-channel irrigation canal network. While no definitive evidence of a matching outbreak strain was found in the on-farm soil and water samples as tested by pulsed-field gel electrophoresis (PFGE), a probable cause of produce contamination has been suggested (Gelting and Baloch, 2013). The Gila River watershed where the implicated farm is located experienced higher than average precipitation and rainfall intensity in the growing season of the outbreak (Crawford and Baloch, 2010), creating a high runoff potential (Cann et al., 2013). Agricultural and urban runoff can increase the microbial load of impacted water sources (Lothrop et al., 2018). Importantly, an adjacent RV park had a septic system with multiple drainage fields on soil that was not suitable for septic adsorption. It is possible that the runoff pathways of these drainage fields were exacerbated by the recent heavy precipitation, thus moving septic waste through the environment (Gelting and Baloch, 2013).

Another recent foodborne outbreak with strong links to contaminated irrigation water is the romaine lettuce outbreak in 2018. There were 210 confirmed *E. coli* O157:H7 infections, resulting in 96 hospitalisations, 27 cases of HUS, and five deaths across 36 states (CDC, 2018). This was the largest *E. coli* outbreak in the USA since the bagged spinach outbreak of 2006 (CDC, 2006). The FDA's environmental assessment detected *E. coli* O157:H7 in an open irrigation canal in the Yuma growing region a few months after the outbreak began. However, this was not an identical match. CDC laboratory testing identified the outbreak strain in canal water, closely related genetically via WGS. This pathogen was not detected in

any other environmental samples. Water contaminated with the pathogen may have come into contact with the lettuce directly, or indirectly through the application of agricultural chemicals. There are several possibilities for how the irrigation canal may have become contaminated (Beach, 2018). Adjacent to the implicated irrigation canal was a large concentrated animal feeding operation (CAFO). Ruminants are reservoirs of *E. coli* O157:H7 and it has been found that 30% of feedlot cattle shed the pathogen (Callaway et al. 2009); however, the limited number of samples taken from this CAFO did not reveal the outbreak strain. Alternatively, there could be a groundwater influence. Groundwater is pumped directly into the canal at two points. Also, sections of the irrigation canal are unlined, allowing possible groundwater recharge with contaminated water. Importantly, the flow of groundwater in this catchment aligns the CAFO with the unlined sections of the canal, despite the CAFO being 'upstream' from a surface water perspective. The time it took for the outbreak to be linked to specific farms in Yuma County may have meant that further evidence linking the lettuce to contaminated water was missed. This also highlights the importance of shipping and distribution records in tracing foodborne disease sources.

Other examples of fresh produce outbreaks linked to irrigation water are reported in Table 1.1.

Table 1.1: Examples of fresh produce outbreaks linked to irrigation water

Outbreak	**Causative agent**	**Evidence implicating irrigation water**	**Reference(s)**
Iran, 2005 1118 cases	*Vibrio cholera* Inaba on vegetables	outbreak agent was detected in irrigation water and vegetable samples	Jonaidi Jafari et al., 2007
USA, 2008 32 cases	*Salmonella* Saintpaul in salsa	outbreak strain in an agricultural irrigation water source and on serrano peppers found in an associated field	Mody et al., 2011
Italy, 2014 206 cases	*Salmonella* Typhimurium	outbreak strain was detected in nine effluent samples, eleven surface water samples, and one irrigation water sample	Cito et al., 2014
USA, 2002 and 2005 510 and 72 cases	Salmonella Newport on tomatoes	irrigation pond was sampled and found to be positive for the outbreak strain	Greene et al., 2008
UK, 2015 47 cases	*E. coli* O157:H7 PT8 in salad	irrigation samples were negative for the outbreak strain (ovine origin) when tested but farms used untreated irrigation water from open ponds or rivers	Mikhail et al., 2018
England, 2013 28 cases	*E. coli* O157 PT2 on watercress	isolated *E. coli* O157 PT2 from exit irrigation water, identical to the outbreak	Jenkins et al., 2015

		strain according to whole genome sequencing	
Sweden, 2005 108 cases	*E. coli* O157 on lettuce	identical strain to the outbreak strain was isolated from cattle upstream; irrigation water and one sample from the irrigation intake pipe were positive for the stx2 gene by PCR	Söderström et al., 2008

1.3 Current recommendations for irrigation water use

1.3.1 Water sources

The source of irrigation water has major implications for its microbial quality and thus the risk it poses to the health of fresh produce consumers. Surface water is inherently more susceptible to faecal inputs than groundwater (Allende and Monaghan, 2015; Pachepsky et al., 2011), yet groundwater can still be exposed to the same risks through aquifer recharge from contaminated sources, or from contamination via unprotected bore openings. Improper storage or transport of otherwise good quality irrigation water can introduce contaminants and lower the quality (Pagadala et al., 2015), yet prolonged storage may exploit natural inactivation processes, thus mitigating the contamination (Murphy et al., 2010). Additionally, biofilms in irrigation systems could be a source of pathogens such as *L. monocytogenes* (Rodríguez-López et al., 2018).

Irrigation water in Australia is primarily sourced from surface water such as irrigation channels or pipelines, rivers, creeks or lakes, and on-farm dams or tanks (ABS, 2022). Fresh produce growers do not always have multiple water sources available to them, in fact geographic location may mean there is a lone water source available for use. Other source water options may not be financially feasible or environmentally responsible, such as town water.

1.3.2 Irrigation application and timing

The method of irrigation used has significant impacts on the risk of fresh produce contamination (Alum et al., 2011; Choi et al., 2004; Psarras et al., 2014; Solomon et al., 2002). Irrigation method is a critical factor in determining which specific part of a crop is most likely to be contaminated (Alum et al., 2011), and can influence the possibility of a pathogen being internalised (Erickson et al., 2010). Surface drip irrigation will lead to greater pathogen contamination of the plant stem compared to plant roots; at the same time, subsurface drip irrigation can reduce (Alum et al., 2011) or eliminate (Jablasone et al., 2004) the risk of pathogen transfer. If irrigation water directly contacts the edible portion of the crop,

experiments have shown that pathogens contacting the abaxial (lower) side of spinach leaves have higher survival rates and a higher probability of internalisation (Erickson et al., 2010). Internalisation can also be via the roots (Hirneisen et al., 2012) and is an issue as it reduces the effectiveness of post-harvest decontamination efforts. There is some evidence that microbial contaminants and pathogens can be internalised by fresh produce crops (Guo et al., 2002; Miles et al., 2009) and that irrigation significantly increases pathogen persistence on injured plants (Barker-Reid et al., 2009; Harapas et al., 2010). Therefore, irrigation methods that avoid or minimise contact between the irrigation water and the edible portion of a crop are ideal for reducing contamination risk (Allende and Monaghan, 2015).

The persistence of microbial contaminants on produce is impacted by meteorological conditions out of the control of growers. Factors such as air temperature, solar exposure, and rainfall (Belias et al., 2020; Ward et al., 2015) can impact bacterial die-off. The use of withholding periods is one of the strategies used in agriculture to reduce bacterial loads on fresh produce crops and help ensure that the product is safe for human consumption. Warmer temperatures and high humidity promote bacterial growth and survival (Ward et al., 2015), while colder temperatures and sunlight exposure can reduce bacterial populations. Growers exploit these natural decay processes to allow bacterial die-off. The type of produce and the specific bacteria present can also determine the effectiveness of any withholding period between irrigation water application and harvest (Belias et al., 2020).

1.3.3 Water quality requirements for certification

In Australia, the fresh produce industry is mostly self-regulating, relying on stringent, industry-lead certification guidance rather than legislation to achieve excellent food safety. There is some national legislation surrounding water use in the pre-harvest production environment, namely the Primary Production and Processing Standard for Leafy Vegetables (Standard 4.2.8) ("Australia New Zealand Food Standards Code – Standard 4.2.8 – Primary Production and Processing Standard for Leafy Vegetables", 2022) which states that "a primary horticulture producer and a primary horticulture processor must take all reasonable measures to ensure that [water] inputs do not make leafy vegetables unacceptable". Aside from complying with legislation, retailers and quality assurance guidelines may request additional water quality criteria be met by growers to maintain certification.

Freshcare is Australia's largest food safety certification scheme for fresh food producers. For a grower to be certified by Freshcare, irrigation water used within 48 hours of harvest on fresh produce that is usually eaten raw and/or minimally processed, and which does not undergo a subsequent, verified pathogen reduction step, is required to be below 100 colony forming units (cfu) of *E. coli* per 100mL (Freshcare Ltd, 2020). While not legislated, the

Australian and New Zealand Environment and Conservation Council (ANZECC) guidelines (ANZECC & ARMCANZ, 2000) recommend that raw food crops in direct contact with irrigation water have a median concentration of thermotolerant coliforms less than 10 cfu per 100mL, based on a regular monitoring program.

A comparison of a selected group of food safety certification schemes or guidelines that cover fresh produce are presented in Table 1.2. While these guidelines and certification schemes have a common goal of ensuring food safety through water quality, it is notable that each system has a different testing requirement. These discrepancies highlight the varied approaches to water risk management globally, and varied tolerances of risk, but might also create confusion for growers or consumers who are not familiar with the reasons behind the lack of a unified target.

Table 1.2: Microbial limits for irrigation water from guidelines, legislation, and certification schemes

Guideline or scheme	Microbial limits for irrigation water
WHO (global)	<1,000 *E. coli* per 100 mL
FSMA (USA)	<126 *E. coli* per 100 mL (geometric mean)
Freshcare (AUS)	<100 *E. coli* per 100 mL (applied within 48 hours of harvest to the edible portion of a crop usually eaten raw and/or minimally processed)
Red Tractor (UK)	<10 *E. coli* per 100 mL (commercial food crops eaten raw and with a significant risk of pathogen contamination)
	<100 *E. coli* per 100 mL (commercial food crops eaten raw)
GLOBALG.A.P. (global)	Local legislation or default to WHO limit
SQF (global)	None
HARPS (AUS)	None
NZGAP (NZ)	None
ANZECC (AUS, NZ)	<10 thermotolerant coliforms per 100 mL (direct contact with irrigation water)
	<1,000 thermotolerant coliforms per 100 mL (no direct contact with irrigation water)

1.4 Monitoring water quality for food safety

1.4.1 Indicators of water quality

The cost and time needed to detect pathogens in water is prohibitive, which is why microbial indicators are a good strategy for monitoring water quality (Park et al., 2013). Indicators are easy to test for and are generally not pathogenic or harmful themselves. However, an individual microbial water quality indicator can be considered insufficient for predicting the overall safety of fresh produce (Won et al., 2013). No single water quality indicator can reliably assess the total risk of microbial contamination in irrigation water (Brookes et al., 2005), and coliform indicators alone cannot provide conclusive information about all the pathogens found in irrigation water sources (Pachepsky et al., 2014).

Studies exploring the value of indicators have shown weak relationships between specific pathogens and common indicators. An early study found that *Campylobacter* density is not significantly correlated to faecal indicators (Carter et al., 1987), likely because using thermotolerant faecal indicators is not a reliable indicator of the presence or abundance of thermophilic Campylobacter, as they persist differently in the environment (St-Pierre et al., 2009). The presence or absence of *Campylobacter coli* and *Campylobacter jejuni* cannot be correlated to *E. coli* or enterococci as shown by a binary logistic regression model (Ahmed et al., 2010). Also, this pathogen-indicator relationship may differ spatially. A study of a river basin in Canada (Wilkes et al., 2009) demonstrated that relationships between indicator bacteria, pathogens, and protozoan parasites were generally weakly correlated, and dependent on the season and site. Further, indicator bacteria densities were inversely related to *L. monocytogenes* (Wilkes et al., 2009).

Still, surrogates that represent evidence of pathogen risks are useful as a measure of risk. Elevated levels of *E. coli* in a water body still indicate an increased probability of pathogen presence (Holvoet et al., 2014); long-term monitoring of microbial indicators remains a reliable method for assessing the potential degree of pathogenic contamination for a specific water body (Wu et al., 2011). Ultimately, the indicators reveal faecal contamination and thus serious risks to human health, regardless of whether specific pathogens are detected (Wu et al., 2011).

1.4.2 Targeted molecular assays and next-generation advancements

Molecular techniques are being increasingly applied to water microbiology surveillance because they can offer rapid identification of human health risks. Microbial source tracking (MST), also known as faecal source tracking, is a group of methods used to discriminate between human and nonhuman origins of faecal contamination (US EPA, 2005). MST assists in identifying the dominant sources of faecal contamination that cause the deterioration of water quality, which helps focus land and water management activities (Rock et al., 2015). As an example of how MST can be applied to fresh produce food safety, a study used a universal *Bacteroidales* marker to identify potential faecal contamination in fresh produce crops, source and irrigation waters, and the hands of harvesters to identify risks in the production environment (Ravaliya et al., 2014). However, PCR assays, including those as part of MST approaches, cannot distinguish between viable and dead cells. For example, there was no difference in the number of *E. coli* cells quantified between secondary and tertiary treated wastewater in one study (Truchado et al., 2016), despite the knowledge that the additional treatment step would have rendered those cells unviable, and

thus reduced human health risk. The inability to distinguish cell viability leads to challenges in interpreting results in conventional MST studies.

When determining the level of microbial risk associated with different water sources, the recent development of next-generation sequencing (NGS) presents an alternate strategy to conventional, culture-based methods (Cao et al., 2017) and avoids the need to predetermine target bacterial organisms. NGS involves the sequencing of millions of DNA fragments in parallel and so can replace the conventional, culture-dependent approaches to isolating and identifying microbial pathogens. NGS is one of the most important culture-independent approaches for exploring microbial water quality. Garner et al. (2021) have identified key applications of NGS where water quality research has been revolutionised, including taxonomic classification, pathogen detection and functional profiling. This has helped researchers and industry members to address complicated microbiological challenges in the water industry.

Microbial communities in a range of freshwater environments have already been characterised by NGS (Nevers et al., 2020; Roberto et al., 2018; Shang et al., 2022; Shilei et al., 2020), as has the microbiome of a variety of fresh produce commodities (Asaff-Torres et al., 2017; Sequino et al., 2022; Telias et al., 2011). A number of studies have explored the microbial communities of agricultural water sources (Abia et al., 2018; Chen et al., 2018; Staley et al., 2014), but relatively few studies have classified the risk of irrigation water as it relates to fresh produce food safety, as reviewed in Jagadeesan et al. (2019). As yet, there is no standard approach to using NGS for monitoring water quality, which will be needed to overcome inconsistencies in approaches, as well as to facilitate the adoption of NGS technologies in the industry (Garner et al., 2021). Additionally, NGS sequencing is like other molecular techniques in that it cannot distinguish between viable and dead cells.

1.4.3 Forecasting and modelling

The development and use of statistical or mechanistic models to predict microbial contamination in freshwater has a clear benefit to the fresh produce industry. Being able to quantify or qualify irrigation water source contamination between routine testing would give growers increased confidence in the safety of their water sources. This is especially pertinent for growers who only test their water sources periodically.

Predictive models have been developed through employing highly correlated environmental and/or water quality parameters measured *in situ* to anticipate *E. coli* or FIB loads (Farnham and Lall, 2015; Herrig et al., 2015; Islam et al., 2018; Mohammed et al., 2017). Recent research has also highlighted the importance of incorporating physical soil characteristics in

improving the power of models designed to predict FIB (Tousi et al., 2021). Additionally, research using mechanistic models has linked weather variables such as solar radiation, relative humidity, and temperature to the formation of *Escherichia coli* and *Salmonella enterica* persister cells on leaf surfaces (Brandl et al., 2022). These persister cells may resuscitate under favourable conditions and are potential reservoirs for foodborne outbreaks.

Given that there are links between climate variables and FIBs, it is quite possible that climate change may contribute to heightened risk of pathogen contamination (Duchenne-Moutien and Neetoo, 2021). Similarly, land use change and increased population pressure will impact water quality, movement of surface runoff and wildlife, and colocation of grazing and horticulture, which will in turn affect the risk of foodborne outbreaks as a result of either directly consuming water or indirectly through eating fresh produce that has been irrigated with contaminated water (Karp et al., 2015). Ultimately, efforts to predict foodborne risks arising from irrigation water will need to take into account weather, water quality, land use, and future climate.

1.5 Thesis overview

The Australian Research Council Industrial Transformation Training Centre (ARC ITTC) for Food Safety in the Fresh Produce Industry, under which this thesis was produced, was formed to conduct industry-focused research to develop practical solutions to prevent or minimise food safety risks in fresh produce. The Australian fresh produce industry has an excellent reputation for food safety, but the number of disease outbreaks linked to pre-harvest water globally is a cause for concern. The overarching aim of the following thesis was to improve our understanding of the human health risks present in typical irrigation water sources, using a peri-urban study area.

Chapter 2 explores the relationships between environmental and water quality parameters and concentrations of *E. coli* in surface water sources. The aim of this chapter is to investigate which parameters have strong statistical correlations to *E. coli*, and which parameters can be used to predict *E. coli* concentrations and limits. Several previously published studies have utilised highly correlated environmental and/or water quality parameters measured *in situ* to anticipate *E. coli* or FIB loads (Farnham and Lall, 2015; Herrig et al., 2015; Islam et al., 2018; Mohammed et al., 2017), but the AgWater App (University of Arizona, 2015) and the modelling behind it (Rock et al., 2016) is particularly useful because it has been developed to assist fresh produce growers meet their requirements under the Food Safety Modernization Act (FSMA) (*FDA Food Safety Modernization Act*, 2017).

Chapter 3 describes the bacterial community composition and food safety risks in irrigation water sources. This chapter describes the bacterial communities present in the studied samples by generating taxonomic and functional profiles as assigned by the SILVA genomic database (Quast et al., 2013) and inferred by the KEGG genomic database (Qiu, 2013), respectively. This chapter also introduces the concept of an irrigation pathogen potential index of bacterial genera identified in the sequences as an indicator of fresh produce food safety risk.

Chapter 4 builds upon the previous two chapters by examining the intrinsic and extrinsic drivers of the bacterial populations and potential pathogens. The objectives of this research chapter were to explore the alpha and beta diversity of the collected water and sediment samples; calculate connectedness and cohesion measures (Herren and McMahon, 2017) from the bacterial communities; and evaluate the statistical correlation of *E. coli* to the irrigation pathogen potential index or other bacterial community characteristics of the sampled irrigation water sources.

Chapter 5 is an internship report validating an existing molecular assay for the detection of environmental, bloom-forming *E. coli* (Nanayakkara, 2019). The internship was completed at Sydney Water laboratories to meet the requirements of the ARC ITTC Food Safety in the Fresh Produce Industry's collaboration requirements. Environmental, bloom-forming *E. coli* are a relevant concern to both the drinking water and fresh produce industries because they have the potential to confound the assessment of microbial indicators in water, and hence the project was mutually beneficial. The validated protocol was applied to a subset (25%) of the Hawkesbury River samples to determine whether there was evidence of bloom strains within the system.

Chapter 6 includes discussion on the common themes and questions linking the research chapters and provides recommendations and an overall conclusion based on the findings and points raised in this thesis.

2 Environmental and water quality parameters as predictors of *Escherichia coli* in irrigation water sources

2.1 Introduction

In contrast to Australia's well-regulated drinking water sources, water sources for irrigating food crops are typically not closely monitored for their microbial water quality. Under Australia's largest fresh produce food safety certification scheme, Freshcare, growers are required to test the microbial quality of their water supply monthly during the period of use, or annually following four consecutive tests of acceptable microbial quality, regardless of their operation size (Freshcare Ltd, 2020). By contrast, the New South Wales (NSW) Department of Health's Drinking Water Monitoring Program recommends monthly sampling for small drinking water supplies (i.e., less than 100 people served), to greater than six samples per week for large supplies (i.e., greater than 100,000 people served) (NSW Department of Health, 2005). A shortfall in irrigation water surveillance can lead to unknown risks being introduced into the food chain, leaving a large number of consumers vulnerable to potential microbial contamination and foodborne disease. This risk is especially concerning for the cultivation of fresh produce because the food product can be eaten raw and minimally processed (Rajwar et al., 2016). As a result, pre-harvest water quality is one of the first control points for identifying and mitigating microbial contamination in the production system.

Microbial water quality has been traditionally measured by the presence and concentration of faecal indicator bacteria (FIB) such as *Escherichia coli* (*E. coli*) (Truchado et al., 2016). FIB come from the intestines of warm-blooded animals and are generally considered unable to proliferate in aquatic environments (Luo et al., 2011; Tenaillon et al., 2010), thereby indicating recent faecal contamination. Faecal contamination can be a threat to human health due to the association with enteric pathogens (Truchado et al., 2017). As previously proven, FIB can be highly variable in Australian water sources (Daly et al., 2013; Vincent et al., 2022), which leads to uncertainty. The ability to estimate microbial contamination in irrigation water at its the point of use would be advantageous in reducing this uncertainty and ensuring that water sources are fit for purpose. This may be possible through utilising highly correlated environmental and/or water quality parameters measured *in situ* to anticipate *E. coli* or FIB loads (Farnham and Lall, 2015; Herrig et al., 2015; Islam et al., 2018; Mohammed et al., 2017).

This is the principal behind the University of Arizona's 'Ag Water' application (University of Arizona, 2015), which offers real-time irrigation water risk prediction. The application was developed to assist growers to meet their obligations under the FDA Food Safety

Modernization Act produce safety rule (Rock et al., 2019), determining whether a water source is safe to use on fresh produce from a public health standpoint. Utilising recent weather conditions of the growing area and user-specified data of physicochemical water quality related to the in-question water supply, the likelihood of microbial contamination is calculated using statistical modelling (Rock et al., 2016). A comparable tool that enables a quick, realistic evaluation of irrigation water risk would be beneficial to the Australian fresh produce industry.

Safe, quality fresh produce grown in Australia is recognised by Freshcare certification. Freshcare estimates that approximately 70% of Australia's fresh produce is certified under their scheme (Freshcare, unpublished data). For a grower to be certified by Freshcare, irrigation water used on fresh produce that is usually eaten raw and/or minimally processed, and which does not undergo a subsequent, verified pathogen reduction step, is required to be below 100 colony forming units (cfu) of *E. coli* per 100mL (Freshcare Ltd, 2020). To ensure this, growers must sample their water source monthly during the period of use, or annually if previously considered safe (i.e., four consecutive acceptable samples). If this requirement is not met, then irrigators must wait 48 hours after application to harvest.

There is no published evidence that identifies under which environmental and water quality conditions water sources are likely to meet or exceed this specific Freshcare criterion. Given this gap in knowledge, the main objective of the following research was to explore the conditions surrounding increased microbial contamination of irrigation water sources, and to determine whether *E. coli* concentrations and microbial limit exceedances could be predicted using statistical modelling of physicochemical and environmental variables.

2.2 Methods and materials

2.2.1 Hawkesbury dataset

A field sampling study was undertaken in the upper Hawkesbury River area in the northwest of Sydney, New South Wales, Australia (Figure 2.1). Eight sampling sites were selected to represent two different irrigation water source types: on-river (5), and on-farm storage (3). The river sites included two locations above the confluence of the principal tributaries, the Grose and Nepean Rivers, in addition to three sites along the main branch of the Hawkesbury River, which flows north-east towards the Pacific Ocean. The confluence of the Nepean and Grose Rivers is approximately 25 kilometres upstream of the Hawkesbury River sampling site at Punt Road. The three on-farm storages can be considered naturally intermittent.

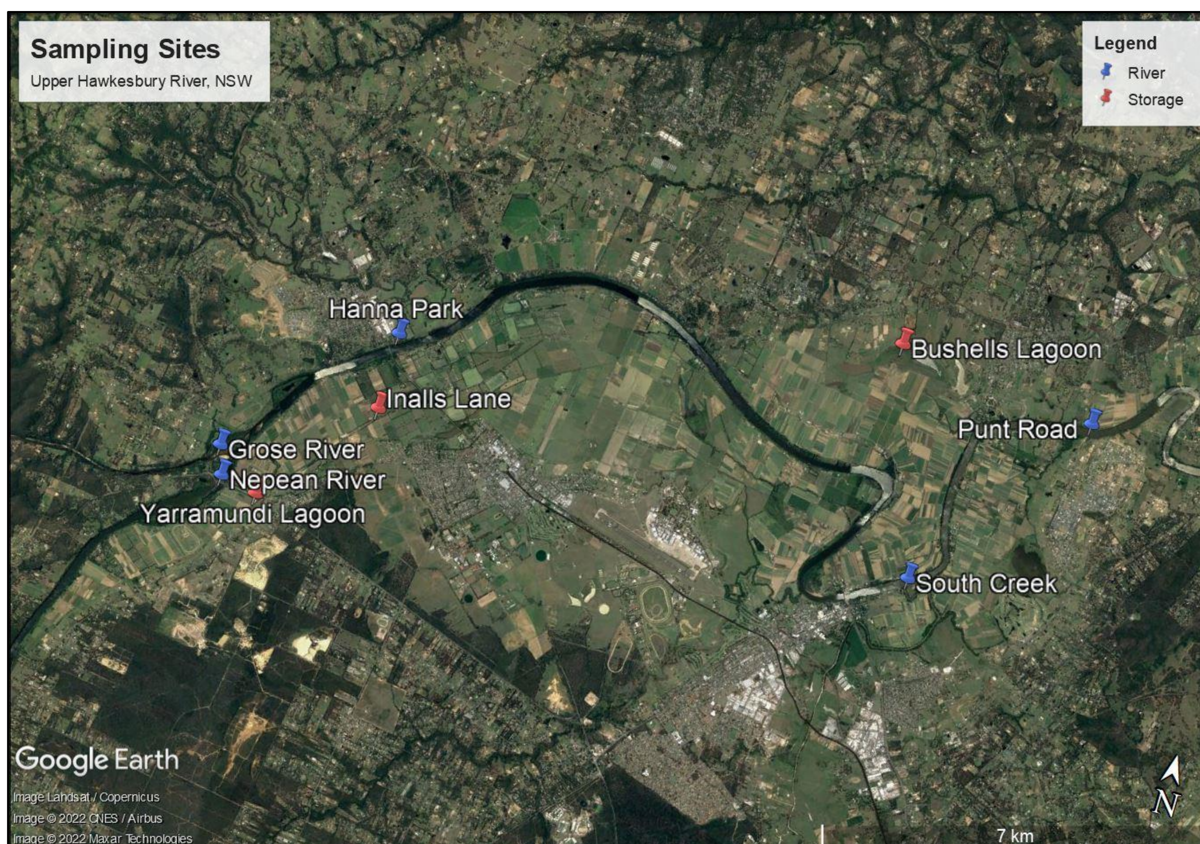


Figure 2.1: Field sampling sites in the upper Hawkesbury River area, New South Wales. River sites (n = 5) are marked blue; on-farm water storages (n = 3) are marked red. The Hawkesbury River begins at the confluence of the Grose and Nepean Rivers (left) and flows north-east towards the Pacific Ocean.

2.2.1.1 Field sampling and testing

Samples were collected monthly from each site for one year to capture seasonal variation. Water (5L from each site) was collected in sterile 1-litre polypropylene bottles using an extendable sampling pole to avoid disturbance along the shoreline. Sediment (150mL from each site) was collected in sterile 50mL tubes from within 0.5m of the shoreline. The exact locations of the sampling were kept consistent, with an exception for Yarramundi Lagoon, where access became unsafe during times of low water levels. The samples were kept on ice for transport to the laboratory, for a maximum of 6 hours.

At the first sampling of each site, a survey was taken of the surrounding area. Primary characteristics were recorded, such as adjacent land use, riparian zone vegetation, presence of macrophytes, topography of the sampling location, and evidence of human activity. Notes were taken on potential contamination sources, including animal, agricultural, industrial, and urban pollutants. At later sampling events, signs of any recent disturbance or change from previous conditions were noted, including evidence of agricultural water pumping.

Water quality parameters were measured *in situ* at each site at the time of sampling with an EXO2 multi-parameter sonde (YSI Incorporated, USA). The sonde was calibrated before use following the manufacturer's instructions, as well as between sampling runs as needed. Parameters included: water temperature (°C), electrical conductivity (mS cm^{-1}), turbidity (NTU), chlorophyll *a* (µg L^{-1}), oxidation-reduction potential (mV), pH, phycocyanin (µg L^{-1}), fluorescent dissolved organic matter (RFU), and dissolved oxygen (%).

Meteorological data was taken from the Australian Bureau of Meteorology (BoM). Daily total rainfall (mm), maximum daily temperature (°C), and daily global solar exposure (MJ m^{-2}) were recorded at the Richmond RAAF base meteorological station, BoM station #67105. This station was chosen as it was central to all sampling sites, being no more than 8 kilometres from the furthest site. Discounted rainfall, temperature, and solar exposure were calculated from this data (Wang et al., 2011) to allow the evaluation of antecedent weather affects. The discount function is essentially an exponential smoothing of the previous data that gives more weighting to the most recent recordings (Wang et al., 2011).

2.2.1.2 Laboratory analysis

Following the first sampling of the sites, the underlying sediment of each study site was classified by particle size distribution using a hydrometer; sand, silt and clay fractions were calculated according to the procedure in the Supplementary Material (page 105).

Following each sampling event, alkalinity (as $CaCO_3$ mg L^{-1}) was determined using Hach Method #8203 (Hach Company, 2018).

Escherichia coli (*E. coli*) was enumerated for each site using the Colilert Quanti-Tray/2000 system (IDEXX Laboratories, USA). For the water samples, 100mL of water was processed according to the manufacturer's instructions. For the sediment samples, 1g of sediment from each site was combined with 100mL of sterile water, which was processed as above. The trays were sealed and incubated at 37°C for 24 hours. Following incubation, generic *E. coli* was quantified with the use of a long-wave ultraviolet (UV) handheld light. Wells that appeared yellow under ambient light and fluoresced blue under UV light were counted as *E. coli*. Wells that appeared yellow or fluorescent but were initially unable to be regarded as positive when compared to the standard were incubated for up to an added 4 hours before being recounted. The most probable number (MPN) of *E. coli* per 100mL was obtained via the conversion tables provided with the Colilert system.

2.2.2 Greater Sydney dataset

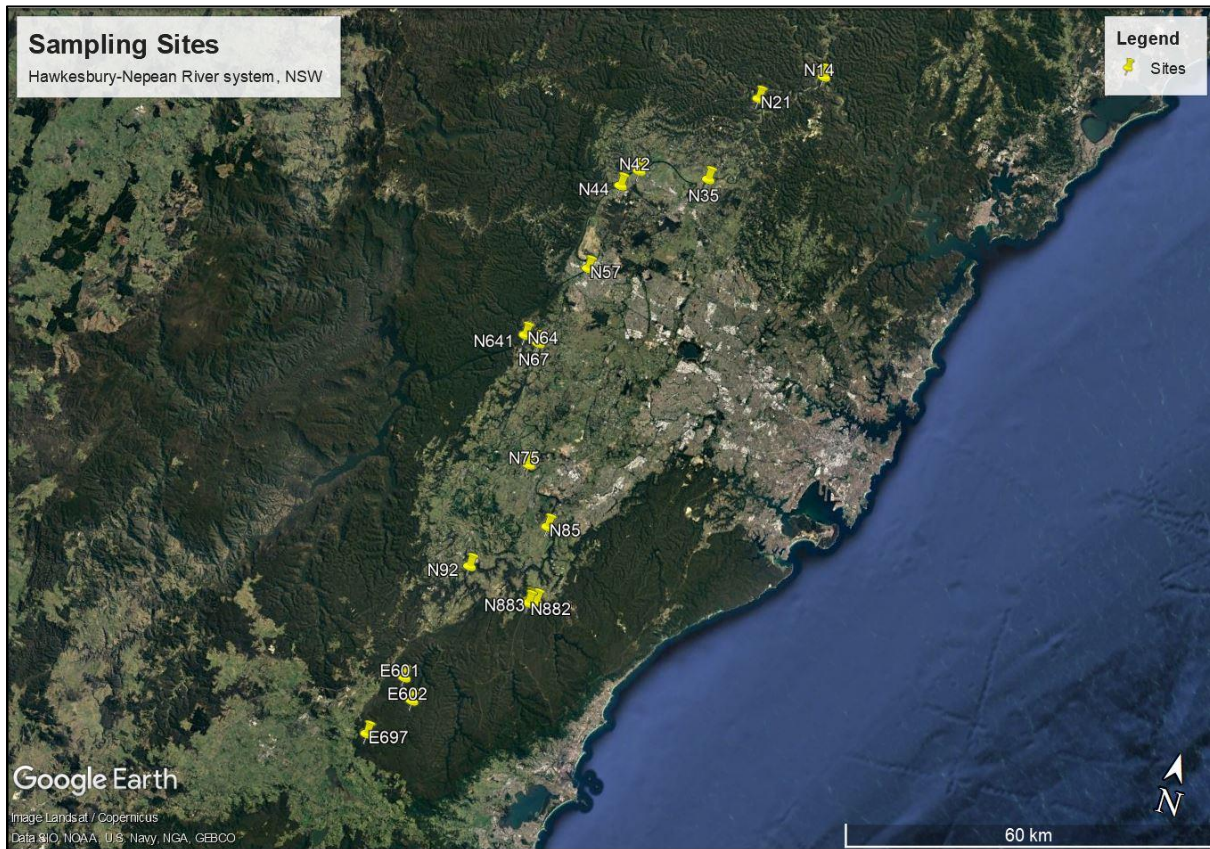


Figure 2.2: WaterNSW sampling sites in the Hawkesbury-Nepean River system, Greater Sydney, NSW. The Nepean River flows north and becomes the Hawkesbury River at its confluence with the Grose River in north-west Sydney.

Water quality data for a length of the Hawkesbury-Nepean River system within the Greater Sydney area (Figure 2.2) was obtained from WaterNSW. Parameters included: water temperature (°C), electrical conductivity (µS cm^{-1}), turbidity (NTU), alkalinity (as $CaCO_3$ mg L^{-1}), chlorophyll *a* (µg L^{-1}), pH, total nitrogen (mg L^{-1}), total phosphorus (mg L^{-1}), total suspended solids (mg L^{-1}), and dissolved oxygen (%). Climate data from the same period was acquired from SILO (Jeffrey et al., 2001), a Queensland Government database based on observations from the BoM's Australian Data Archive for Meteorology (ADAM). Daily total rainfall (mm), maximum and minimum daily temperatures (°C), and daily global solar exposure (MJ m^{-2}) were taken from the closest meteorological station to each site. Discounted climate variables were also calculated from this data (Wang et al., 2011).

2.2.3 Statistical analysis

All of the following statistical analyses were performed in R (R Core Team, 2021).

2.2.3.1 Hawkesbury dataset

Data from the Hawkesbury field sampling study was used to explore general the statistical relationship between the *E. coli* concentrations of the water column to those of the underlying sediment, and to physicochemical and environmental parameters. The IDEXX Colilert system has a quantification limit of 1 MPN per 100mL, causing negative samples to be reported as “<1.0”. The Quanti-Tray/2000s also have an upper quantification limit of 2419.6 MPN per tray, causing quantities above this to be reported as “>2419.6”. This creates what is known as interval censoring in the *E. coli* concentration data and therefore it is valid to use statistical techniques for censored data. The ‘survival’ (Therneau, 2022) and ‘NADA2’ (Julian and Helsel, 2022) packages in R were used for this purpose.

To compare the *E. coli* concentrations in the water column to the concentrations in the underlying sediment, the Akritas-Theil-Sen (ATS) line (Akritas et al., 1995) for censored data was calculated. To compare the *E. coli* concentrations in the water column between sites, seasons, and soil types, the Peto-Peto test (Peto and Peto, 1972) of differences in cumulative distribution functions between groups was used. For significant ($P < 0.05$) results, pairwise multiple comparisons, adjusted using the Benjamini-Hochberg false-discovery rate (Benjamini and Hochberg, 1995), were explored.

To explore the relationship between the concentration of *E. coli* in the water column and the water quality and environmental variables, a generalised additive model (GAM) was fitted using the ‘mgcv’ package (Wood, 2011). Pairwise concurvity between the explanatory variables was checked prior to selecting the final model; concurvity estimates over 0.8 resulted in the less-significant variable being removed for this analysis. To optimise the model, smooth functions with *k* values set to limit basis dimension were fitted.

2.2.3.2 Greater Sydney dataset

The dataset from WaterNSW was used to evaluate the predictability of *E. coli* concentrations in on-river water sources based on physicochemical and environmental variables. This dataset did not contain censored values. Pairwise correlations were calculated to examine the variables for redundancy prior to modelling. Using the ‘randomForest’ package (Liaw and Wiener, 2002), random forest models were constructed for two dependant variables: *E. coli* concentration as a continuous variable (i.e., regression), and *E. coli* concentration as a binary variable (i.e., classification). The binary variable was derived from *E. coli* measurements being within or exceeding the <100 cfu per 100mL requirement for Freshcare certification. The random forest models were improved by adjusting *m* (the number of variables tried at each split) to obtain the lowest possible out of bag (OOB) error. To calibrate and validate the models, the data was split 75:25 and model performance statistics

were evaluated. The model performance statistics for the classification model are detailed in Figure 2.3.

True positive (TP): predicted positive, observed positive

True negative (TN): predicted negative, observed negative

False positive (FP): predicted positive, observed negative

False negative (FN): predicted negative, observed positive

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Specificity: $\frac{TN}{TN+FP}$

Precision: $\frac{TP}{TP+FP}$

Sensitivity: $\frac{TP}{TP+FN}$

False positive rate (FPR): $\frac{FP}{TN+FP}$

False negative rate (FNR): $\frac{FN}{TP+FN}$

Figure 2.3: Definitions and equations of the statistics used to evaluate the performance of the random forest classification model (i.e., *E. coli* concentrations exceeding the Freshcare irrigation water requirement).

2.3 Results

2.3.1 Hawkesbury dataset

2.3.1.1 Climate variables

During the 12-month period of field sampling, the weather varied seasonally (Figure 2.4). The upper Hawkesbury River area received most of its annual rainfall in summer, i.e., December to February (Figure 2.4A) but did receive several smaller (<20mm per day) rainfall events during the winter months of sampling, i.e., June to August. The maximum daily temperature ranged from approximately 40°C in the peak of summer to 15°C during winter (Figure 2.4B). Global solar exposure was highest during summer (Figure 2.4C) but fluctuated year-round.

2.3.1.2 Site characteristics

There were diverse land uses and potential contamination sources observed in the study area (Table 2.1). Agriculture, recreation, and industrial land uses were observed in the immediate vicinity of the water sampling sites, and there was evidence of birds, dogs, horses, and agricultural chemicals. For most of the sites, the topography of the surrounding area (i.e., embankments, slopes) made them vulnerable to contamination from surface run-off. All the sampling sites had a vegetated riparian zone to some extent, but the composition of vegetation types and density varied. There were aquatic macrophytes at Bushells Lagoon and Hanna Park. The particle size analysis of the sediments showed that all the on-river sites had an underlying sand soil type; the three on-farm storages differed in soil type, ranging from clay to sandy loam.

In terms of physicochemical water quality (Table 2.2), water temperature, pH and ORP were comparable between all the sampling sites. Bushells Lagoon had notably higher and more variable turbidity (mean 86 NTU; standard deviation 123 NTU) and conductivity (mean 880 μS cm^{-2}; standard deviation 427 μS cm^{-2}) than the other sites. There was a notable contrast between the on-river and the storage sites for several of the water quality variables; chlorophyll *a*, phycocyanin, dissolved organic matter, and alkalinity were generally higher for the on-farm storages.

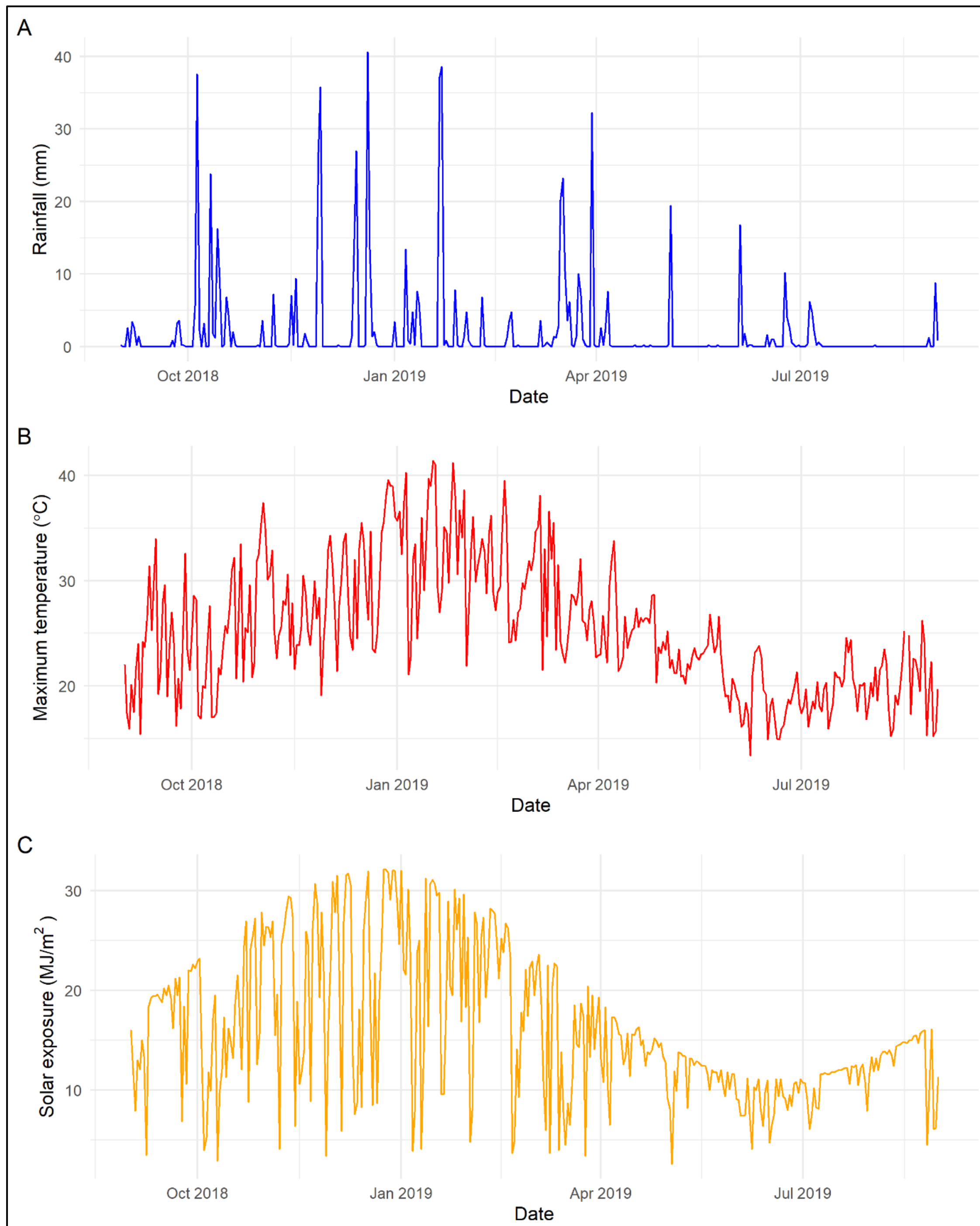


Figure 2.4: A) daily rainfall (mm), B) daily maximum temperature (°C), and C) daily global solar exposure (MJ m^{-2}) at BoM station #67105 (Richmond RAAF) during the period of field sampling, i.e., spring 2018 to winter 2019

Table 2.1: Hawkesbury region field site sampling locations and important site characteristics

Site	Coordinates	Water source	Land use	Potential contamination	Vegetation	Topography	Soil
Bushells Lagoon	-33.563, 150.820	Storage; intermittent	Agriculture	Agricultural chemicals; birds	Grassy and woody riparian zone; macrophytes	Depression	Clay (40% clay; 43% silt; 17% sand)
Grose River	-33.609, 150.697	River; shallow, moderate flow	Heavy mixed-use recreation (primary and secondary contact)	Human recreational activity; dogs; horses	Extensive riparian zone	Downstream of mountainous area	Sand (2% clay; 1% silt; 97% sand)
Hanna Park	-33.584, 150.725	River; deep, slow flow	Mixed-use recreation (secondary contact); nearby agriculture and industry	Human recreational activity; birds; dogs; pesticide use; major road	Riparian zone on one side; grass and sparse woody vegetation on other side; macrophytes	Embankment one side; slightly declined other side	Sand (5% clay; 5% silt; 90% sand)
Inalls Lane	-33.597, 150.725	On-farm dam; pumped water and rainfed	Agriculture	Agricultural chemicals	Sparse riparian zone	Channel on plain	Sandy loam (20% clay; 10% silt; 70% sand)
Nepean River	-33.614, 150.699	River; shallow, moderate flow	Mixed-use recreation (primary contact); nearby agriculture	Human recreational activity; birds; dogs; pesticide use	Riparian zone on one bank; sparse woody vegetation on other	Embankment one side; slightly declined other side	Sand (1% clay; 1% silt; 98% sand)
Punt Road	-33.568, 150.859	Tributary	Mixed-use recreation (secondary contact); nearby agriculture	Human recreational activity; earthworks at top of embankment	Woody riparian zone	Embankments and narrow shoreline	Sand (3% clay; 1% silt; 96% sand)
South Creek	-33.600, 150.833	Tributary, near confluence; interchangeable flow	Heavy mixed-use recreation (primary and secondary contact)	Human recreational activity; birds; dogs; pesticide use	Riparian zone on one bank; grass and sparse woody vegetation on other	Embankments	Sand (6% clay; 6%silt; 88% sand)
Yarramundi Lagoon	-33.613, 150.708	On-farm dam; rainfed	Agriculture	Agricultural chemicals; birds	Very sparse riparian zone	Depression	Clay loam (30% clay; 36% silt; 34% sand)

Table 2.2: Mean water quality conditions at the Hawkesbury region field sites during the period of sampling (standard deviation in brackets). DO = dissolved oxygen; PC = phycocyanin. *The values for the *E. coli* concentrations are Kaplan-Meier estimates of the restricted mean and standard deviation due to the censoring of the data

Site	**Water temp (°C)**	**pH**	**Turbidity (NTU)**	**Cond (µS cm^{-1})**	**DO (%)**	**Chlorophyll *a* (µg L^{-1})**
Bushells Lagoon	21.15 (6.48)	7.39 (0.52)	85.79 (123.06)	879.94 (427.4)	67.64 (35.9)	15.8 (9.11)
Grose River	21.85 (6.58)	7.79 (0.47)	15.23 (22.06)	235.75 (112.73)	106.55 (6.65)	1.65 (1.94)
Hanna Park	21.51 (6.36)	7.75 (0.52)	14.98 (22.88)	341.02 (78.09)	100.65 (10.44)	2.86 (1.06)
Inalls Lane	22.08 (6.59)	7.93 (0.51)	41.45 (30.09)	376.59 (90.73)	101.16 (35.25)	13.87 (10.07)
Nepean River	22.28 (6.4)	8.27 (0.84)	16.44 (21.64)	391.5 (94.94)	115.9 (23.33)	2.42 (0.82)
Punt Road	20.31 (5.35)	7.64 (0.46)	40.09 (33.75)	460.49 (112.49)	90.12 (13.24)	6.94 (4.48)
South Creek	20.78 (5.66)	7.69 (0.44)	54.42 (53.16)	509.88 (131.59)	85.28 (8.11)	5.39 (1.77)
Yarramundi Lagoon	22.13 (6.86)	7.66 (0.56)	59.28 (37.83)	489.71 (129.37)	81.8 (22.72)	10.54 (9.99)
Site	**ORP (mV)**	**PC (µg L^{-1})**	**fDOM (RFU)**	**CaCO3 (mg L^{-1})**	**Water E. coli (MPN 100mL^{-1})***	**Sediment E. coli (MPN g^{-1})***
Bushells Lagoon	157.21 (60.02)	0.72 (0.56)	43.33 (17.84)	92.65 (43.68)	165.80 (288.40)	35.7 (94.9)
Grose River	164.75 (54.2)	0.05 (0.16)	5.19 (4.47)	33.58 (14.34)	88.55 (86.87)	20.4 (38.8)
Hanna Park	166.55 (55.13)	0.09 (0.1)	7.87 (3.1)	52.73 (15.12)	356.80 (312.30)	61.7 (55.1)
Inalls Lane	170.6 (62.21)	0.86 (0.94)	33.02 (13.96)	75.07 (30.62)	117.00 (181.80)	219.8 (391.1)
Nepean River	161.18 (65.72)	0.07 (0.09)	9.16 (2.78)	62.22 (18.58)	28.97 (13.45)	53.7 (89.0)
Punt Road	174.77 (52.15)	0.43 (0.29)	15.73 (6.55)	65.05 (20.86)	94.62 (95.87)	33.6 (39.1)
South Creek	170.14 (63.52)	0.24 (0.14)	14.14 (3.48)	69.42 (23.77)	114.70 (131.60)	209.6 (355.4)
Yarramundi Lagoon	172.43 (63.49)	0.38 (0.37)	35.52 (5.53)	99.52 (28.46)	264.30 (473.90)	523.9 (705.3)

2.3.1.3 Comparison of sediment and water

The ATS line (intercept 3.58; slope 0.23; Kendall's tau 0.28) was used to compare the concentrations of *E. coli* in the water column to the concentrations in the sediment (Figure 2.5). Kendall's tau is a measure of the non-parametric relationship between two ranked variables (Kendall, 1938). The P value for the test that tau and the slope both equal zero is < 0.01, meaning there is a significant positive relationship between the *E. coli* in the water and in the sediment.

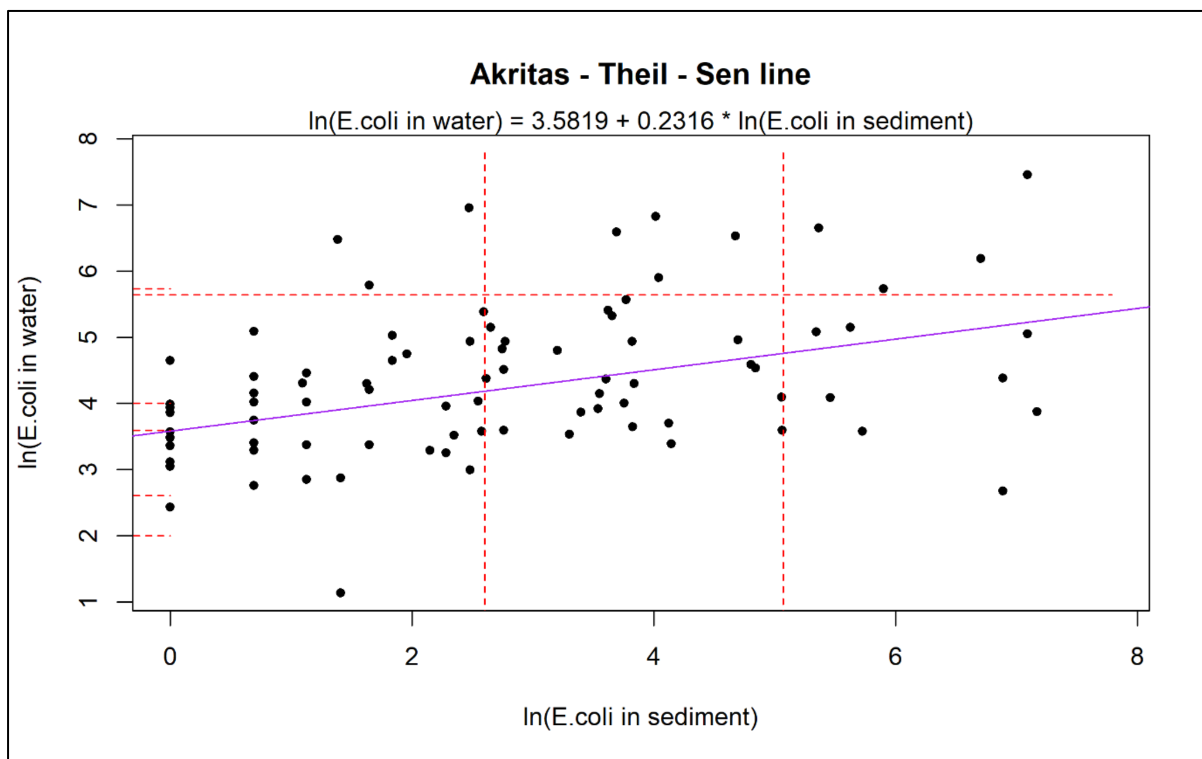


Figure 2.5: A scatterplot of log-transformed *E. coli* concentrations in the sediment (x-axis) and the water (y-axis) with the fitted Akritas-Theil-Sen (ATS) trendline in purple (intercept 3.58; slope 0.23)

2.3.1.4 E. coli contamination

The Peto-Peto nonparametric tests revealed that there were significant differences (Chi-squared value = 40.07; 7 degrees of freedom; P < 0.01) in the water *E. coli* concentrations between sampling sites, but not between different seasons, soil types, or rivers compared to on-farm dams.

Pairwise comparisons (Table 2.3) showed that the Hawkesbury River at Hanna Park had the highest mean *E. coli* concentration in the water column (356.8 MPN 100mL^{-1}) over the period of sampling, but it was not significantly higher than the next highest mean concentration at Yarramundi Lagoon (264.3 MPN 100mL^{-1}). Similarly, the Nepean River just before the confluence with Grose River had the lowest mean *E. coli* concentration (28.97

MPN 100mL^{-1}), but it was not significantly lower than the next lowest mean concentration at Punt Road (94.62 MPN 100mL^{-1}).

Table 2.3: Pairwise comparisons (Peto-Peto test) of *E. coli* concentrations in the water (* indicates significant differences, P < 0.05)

	Bushells Lagoon	Grose River	Hanna Park	Inalls Lane	Nepean River	Punt Road	South Creek
Grose River	ns						
Hanna Park	*	*					
Inalls Lane	ns	ns	*				
Nepean River	*	*	*	*			
Punt Road	ns	ns	*	ns	*		
South Creek	ns	ns	*	ns	ns	ns	
Yarramundi Lagoon	ns	ns	ns	ns	*	ns	ns

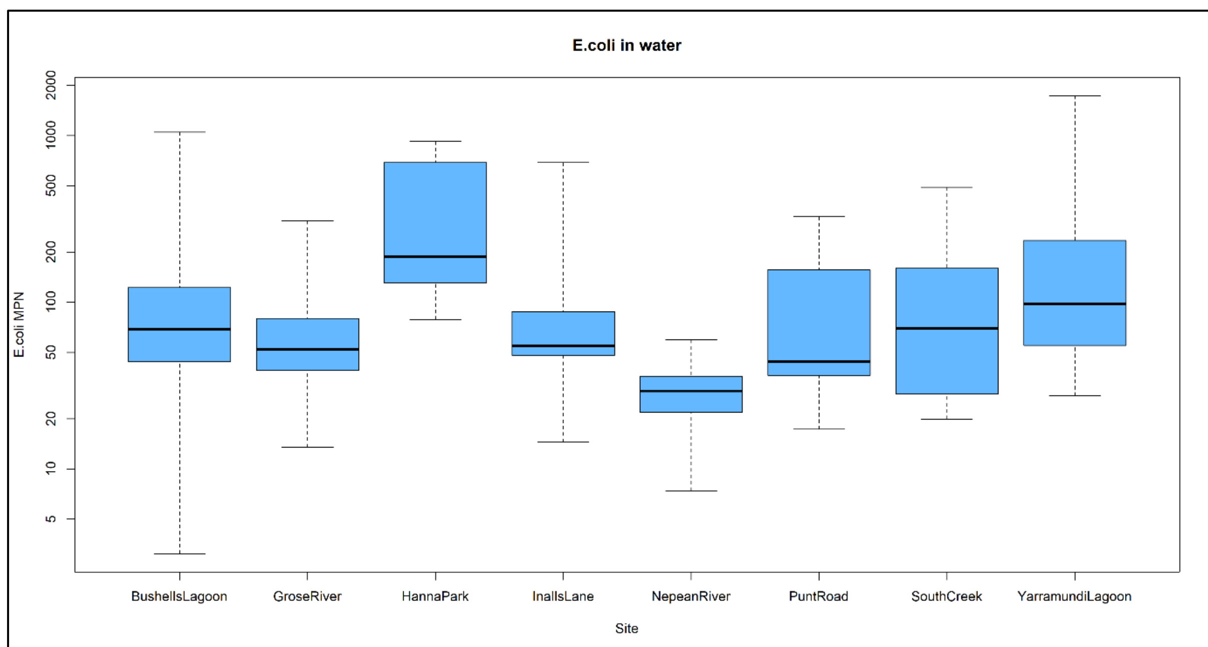


Figure 2.6: Boxplots of *E. coli* concentrations (MPN per 100mL) in the water samples (n = 96) by site.

While the *E coli* concentrations did not vary significantly by season in the water samples (P > 0.05), there was a significant seasonal trend in the sediment samples (Chi-Squared value = 16.23; 3 degrees of freedom; P < 0.01) (Figure 2.7). Concentrations in autumn (mean 324.7 MPN g^{-1}) were significantly higher than in spring (32.3 MPN g^{-1}; P = 0.017) and winter (17.3 MPN g^{-1}; P < 0.01), but not significantly higher than in summer (59.1 MPN g^{-1}; P > 0.05).

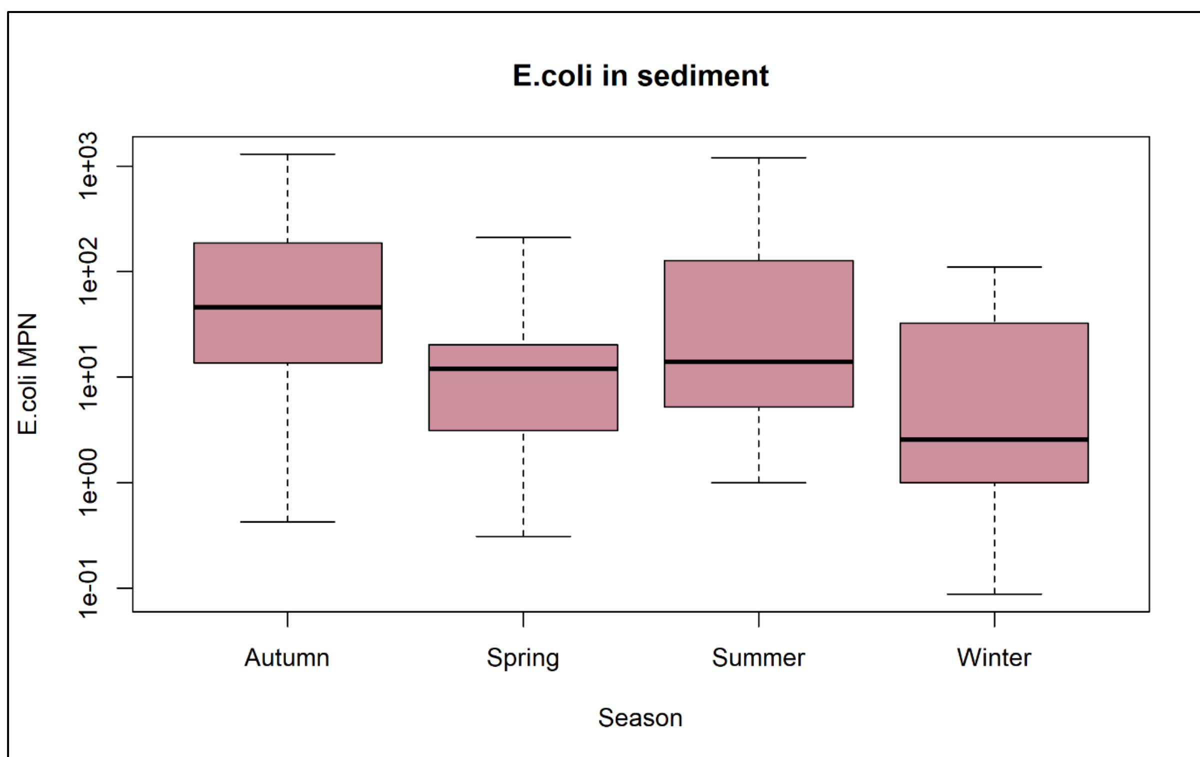


Figure 2.7: Boxplots of *E. coli* concentrations (MPN per 1g) in the sediment samples (n = 94) by season.

2.3.1.5 Generalised additive model (GAM)

The relationship between the concentration of *E. coli* in the water column and the water quality and environmental variables was described by using a generalised additive model (GAM). Due to missingness in the physicochemical metadata, n = 86. The GAM that explained the most deviance (62.9%, Table 2.4) had an adjusted R^2 value of 0.492 and included smooth terms for discounted daily rainfall (DF 0.99), discounted daily maximum temperature (DF 0.5), longitude, conductivity, dissolved oxygen, and turbidity, of which all were significant predictors of *E. coli* ($P < 0.05$), except for turbidity ($P > 0.05$). Linear terms for fDOM, alkalinity, chlorophyll *a*, phycocyanin, pH, and ORP were also included in the GAM, of which fDOM and chlorophyll *a* were significant ($P < 0.05$).

The combined daily rainfall and maximum temperature term was also highly significant ($F = 6.16$; edf = 2; $P < 0.01$). As shown by the spline, *E. coli* concentration increased as maximum temperature and daily rainfall increased (Figure 2.8).

The smooth term for longitude (Figure 2.9) had 5.68 estimated degrees of freedom and was therefore a more complex term. The term had less error at the lower and higher longitude values, reflecting the locations of the sampling sites within the study area (Figure 2.1). In general terms, it appeared that the *E. coli* concentrations were modelled to increase with longitude, i.e., downstream.

Table 2.4: Coefficients for the generalised additive model (GAM) of the concentration of *E. coli* in the water column (n = 86). The GAM explained 62.9% of deviance and had an adjusted R^2 value of 0.492. (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown)

Parametric coefficients	**Estimate**	**Std. Error**	**t value**	**P**	**Sig.**
(Intercept)	0.258	1.048	0.247	0.806	
fDOM	-0.019	0.006	-3.010	0.004	**
Alkalinity	0.003	0.003	1.174	0.245	
Chlorophyll *a*	0.023	0.011	2.110	0.039	*
PC	-0.233	0.160	-1.456	0.151	
pH	0.186	0.123	1.512	0.136	
ORP	0.002	0.001	1.449	0.152	
Approximate significance of smooth terms	**Est. df**	**Ref. df**	**F value**	**P**	**Sig.**
Rain (DF 0.99), Max temp. (DF 0.5)	2.00	2	6.161	3.63E-03	**
Longitude	5.68	6	6.717	7.04E-06	***
Conductivity	3.68	4	6.042	1.29E-04	***
DO	2.14	3	2.098	0.044	*
Turbidity	3.47	4	1.623	0.122	

The smooth term for conductivity (Figure 2.10) had 3.68 estimated degrees of freedom. Most of the conductivity values existed in the range of 200 to 800 mS cm^{-2}, where the low point of *E. coli* concentrations was at approximately 500 mS cm^{-2}, with increases on either side. There were scarce conductivity values above 800 mS cm^{-2}, and therefore the model error increased, but the general trend suggested a negative trend.

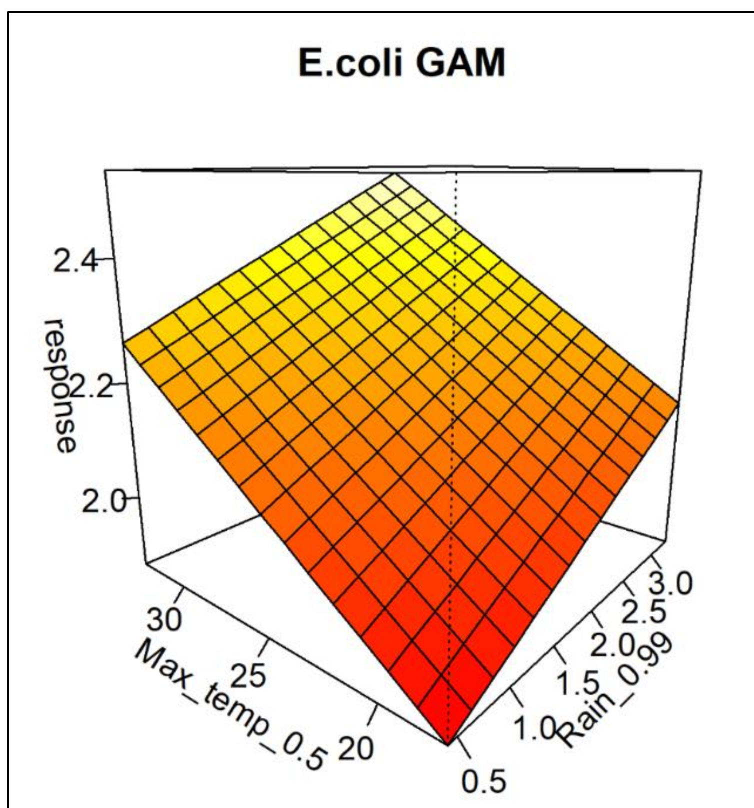


Figure 2.8: The response of log10 *E. coli* concentration (z-axis) to daily rainfall (DF 0.99) (x-axis) and maximum temperature (DF 0.5) (y-axis), as modelled by the GAM.

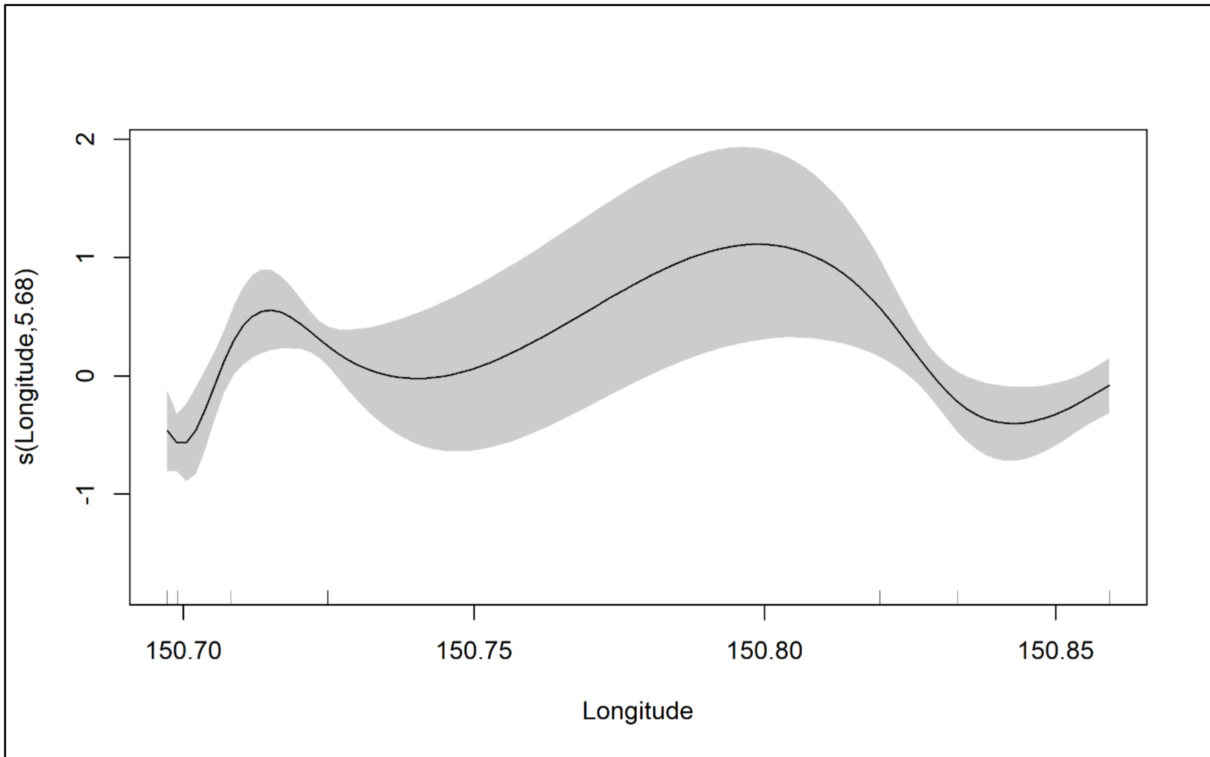


Figure 2.9: A smooth term plot for longitude (edf = 5.86), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis.

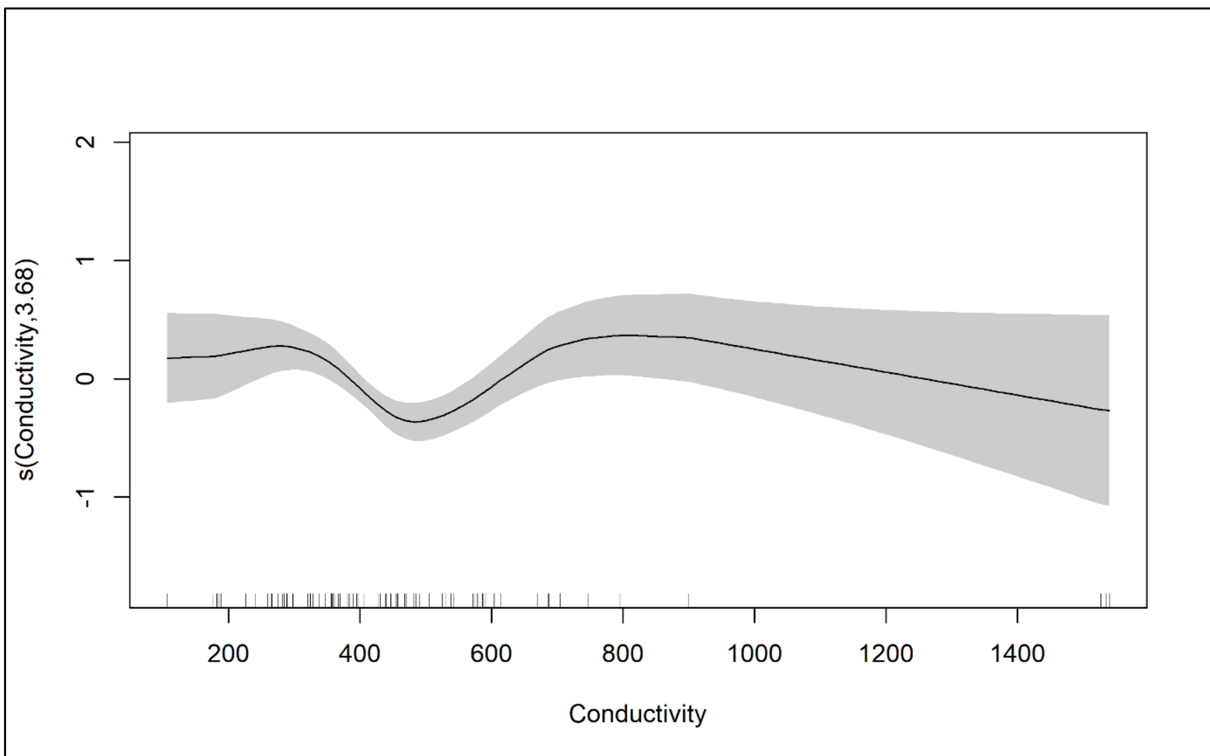


Figure 2.10: A smooth term plot for electrical conductivity (edf = 3.68), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis.

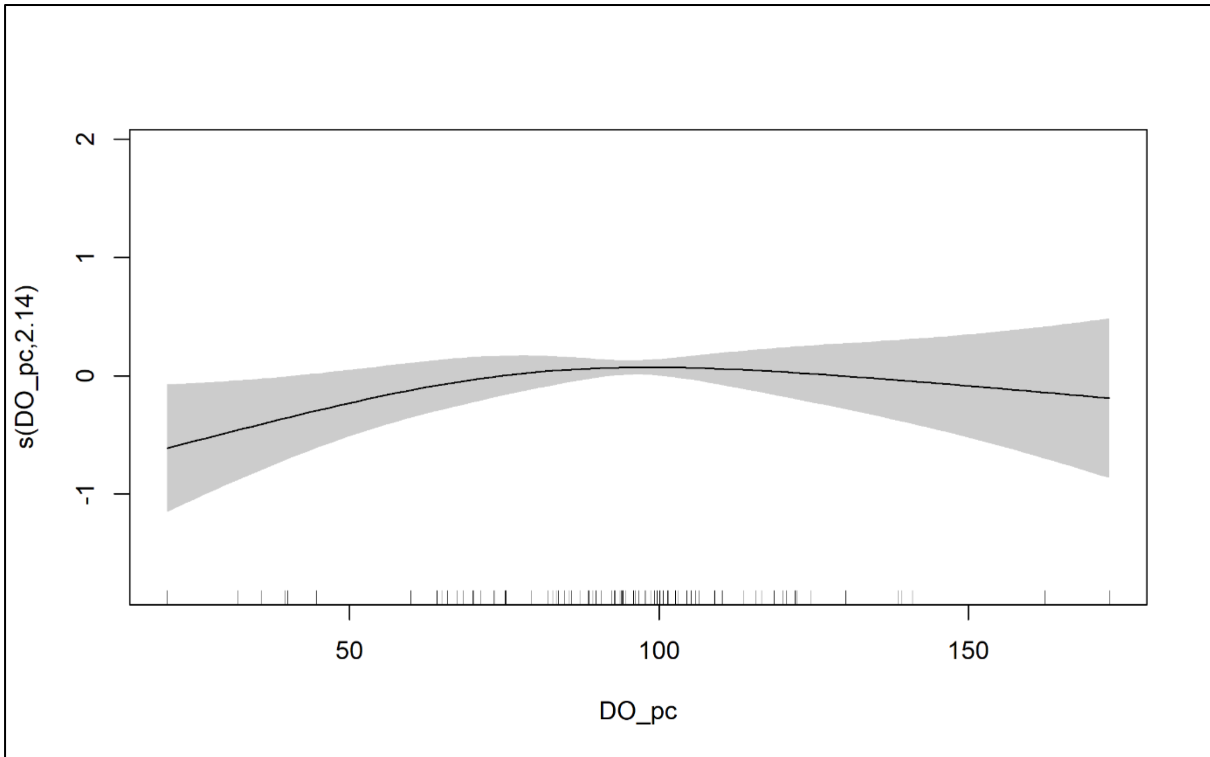


Figure 2.11: A smooth term plot for dissolved oxygen (edf = 2.14), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis.

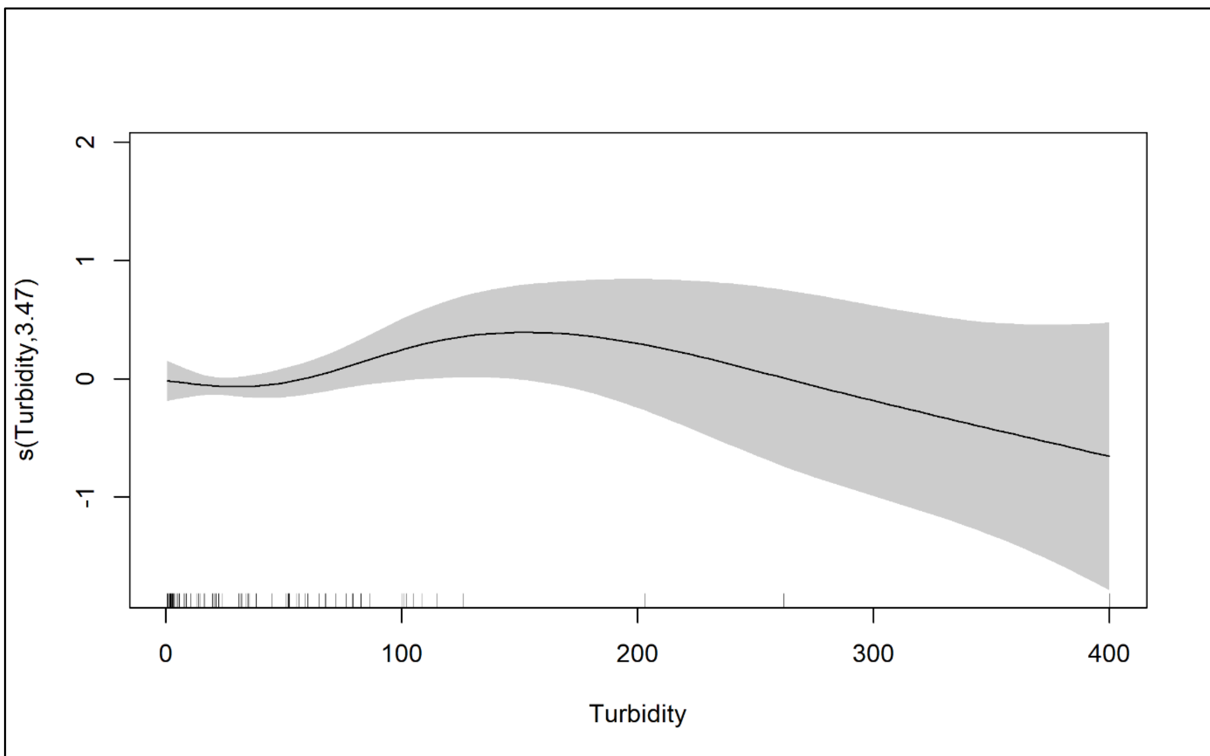


Figure 2.12: A smooth term plot for turbidity (edf = 3.47), as modelled by the GAM. The y-axis shows the log-transformed *E. coli* MPN, the shaded area shows the standard error of the prediction, and the observed values are indicated by tick marks along the y-axis.

The smooth term for DO (Figure 2.11) had 2.14 estimated degrees of freedom. The DO values sat around 100%, where the model error was lowest and where *E. coli* concentrations were highest. *E. coli* gradually decreased outside of the 60 – 140 % DO range.

The smooth term for turbidity (Figure 2.12) had 3.47 estimated degrees of freedom. The model error increased with the turbidity values due to scarcity of data above the 100 NTU range. The smooth term suggests that turbidity values between 0 and 50 do not influence *E. coli*, but that *E. coli* increases between 50 and 100 NTU of turbidity. The GAM plot suggests that turbidity is limiting to *E. coli* at values above 150 NTU, but due to the scarcity of data and large model error, this cannot be interpreted with much confidence.

2.3.2 Greater Sydney dataset

2.3.2.1 Data summary

The Greater Sydney dataset (Figure 2.2) consisted of 1537 data points from 18 water quality stations, covering the period from February 2012 to November 2017. Summary statistics for the water quality variables are shown in Table 2.5.

Table 2.5: Water quality and climate summary statistics (median, mean, standard deviation) for the Greater Sydney dataset

Variable	**Median**	**Mean**	**Stand. Dev.**
E. coli (orgs 100mL^{-1})	19	134.90	895.90
Alkalinity (CaCO3 mg L^{-1})	31	32.90	24.30
Chlorophyll *a* (µg L^{-1})	4.1	6.71	8.08
Conductivity (mS cm^{-1})	0.20	0.32	1.03
Dissolved oxygen (DO) (% sat)	99.52	97.40	12.05
Total nitrogen (TN) (mg L^{-1})	0.42	0.45	0.25
pH	7.4	7.41	0.46
Total phosphorus (TP) (mg L^{-1})	0.01	0.02	0.02
Total suspended solids (TSS) (mg L^{-1})	3	4.38	5.71
Turbidity (NTU)	3.62	6.33	8.84
Water temperature (°C)	18.50	17.9	5.46
Maximum temperature (°C)	22.92	22.69	4.94
Minimum temperature (°C)	11.10	11.06	4.69
Solar exposure (mJ m^{-2})	16.78	16.11	4.55
Daily rainfall (mm)	0.66	3.55	6.74

Model variables with pairwise correlations above 0.7 were considered superfluous. Ambient temperatures (maximum and minimum), water temperature, and solar radiation were all highly positively correlated. Therefore, the weather variables selected for model development included antecedent daily rainfall with a discount factor of 0.75, and antecedent minimum daily temperature with a discount factor of 0.5, as these discount factors lead to

better performing models. Additionally, turbidity and TSS were highly positively correlated, so turbidity was selected for inclusion.

2.3.2.2 Regression model

The random forest model of continuous *E. coli* (with a log10 transformation) with the smallest OOB error tried 9 variables at each split and explained 57.5% of the training data variation. The model had a calibration accuracy of 0.07 (R^2) and Lin's concordence coefficient correlation of 0.10. The variable importance of the parameters is shown in Figure 2.13. Latitude, rainfall, and turbidity were the three most important variables in the model in terms of mean squared error and node purity.

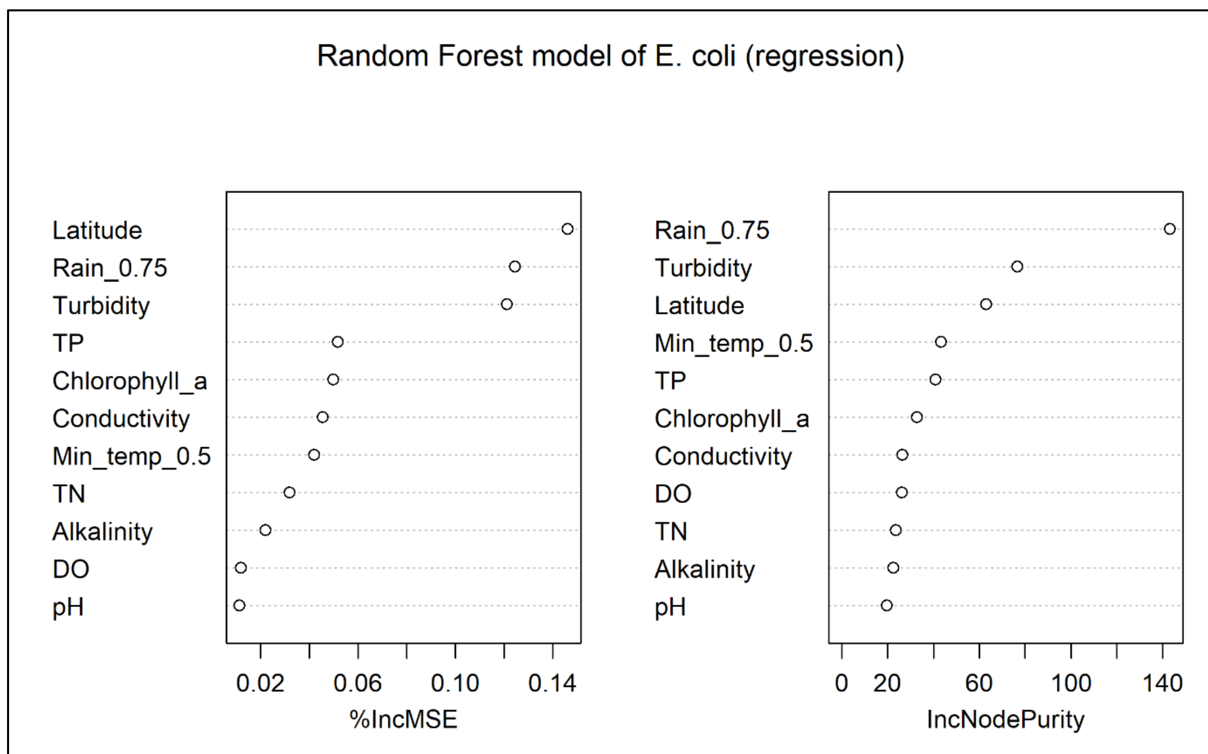


Figure 2.13: Variable importance plots for the regression random forest model predicting *E. coli* concentrations. The left plot shows the percent increase in the mean square error (MSE) when the term was removed from the model. The right plot shows the increase in node purity when the term was included in the model.

Rainfall and turbdity had similar modelled relationships with *E. coli*, as shown in their partial dependence plots (Figure 2.14); as rainfall and turbidity increased, so too did modelled *E. coli*, before plateuing at approximately 20mm and 30 NTU, respectively. The modelled relationship between *E. coli* and latitude was more complex and highly variable; a series of sharp peaks and dips.

The calibrated model was used to predict on the withheld data. The random forest model had a validation accuracy of 0.15 (R^2) and Lin's concordence coefficient correlation of 0.18. The plot of oberved and predicted values (Figure 2.15) showed that the model predicted a

majority of *E. coli* concentrations fairly well, but that a small number of higher *E. coli* concentrations were underestimated by 5 to 10 times their real value.

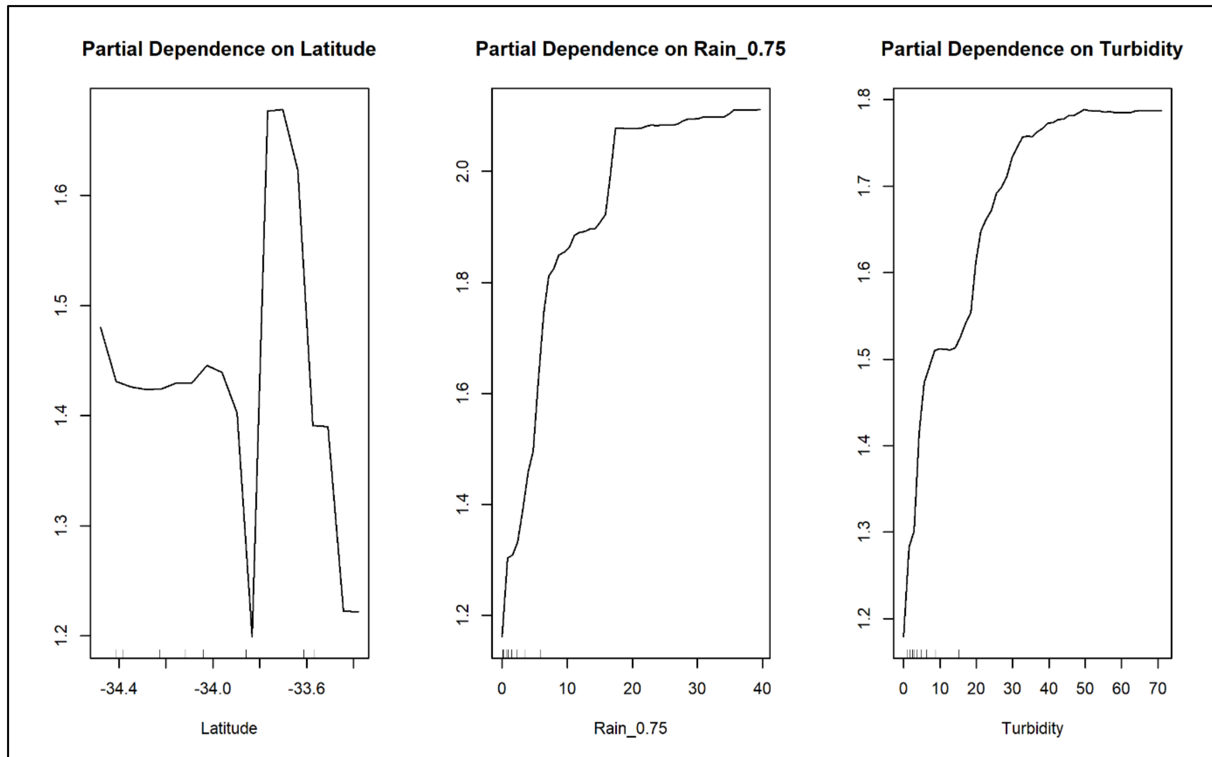


Figure 2.14: Partial (marginal) dependence plots of the regression random forest model for latitude, rainfall (DF 0.75), and turbidity. The y-axis is the modelled response of log10 *E. coli* MPN per 100mL. The observed values used in the model calibration are indicated by tick marks along the y-axis.

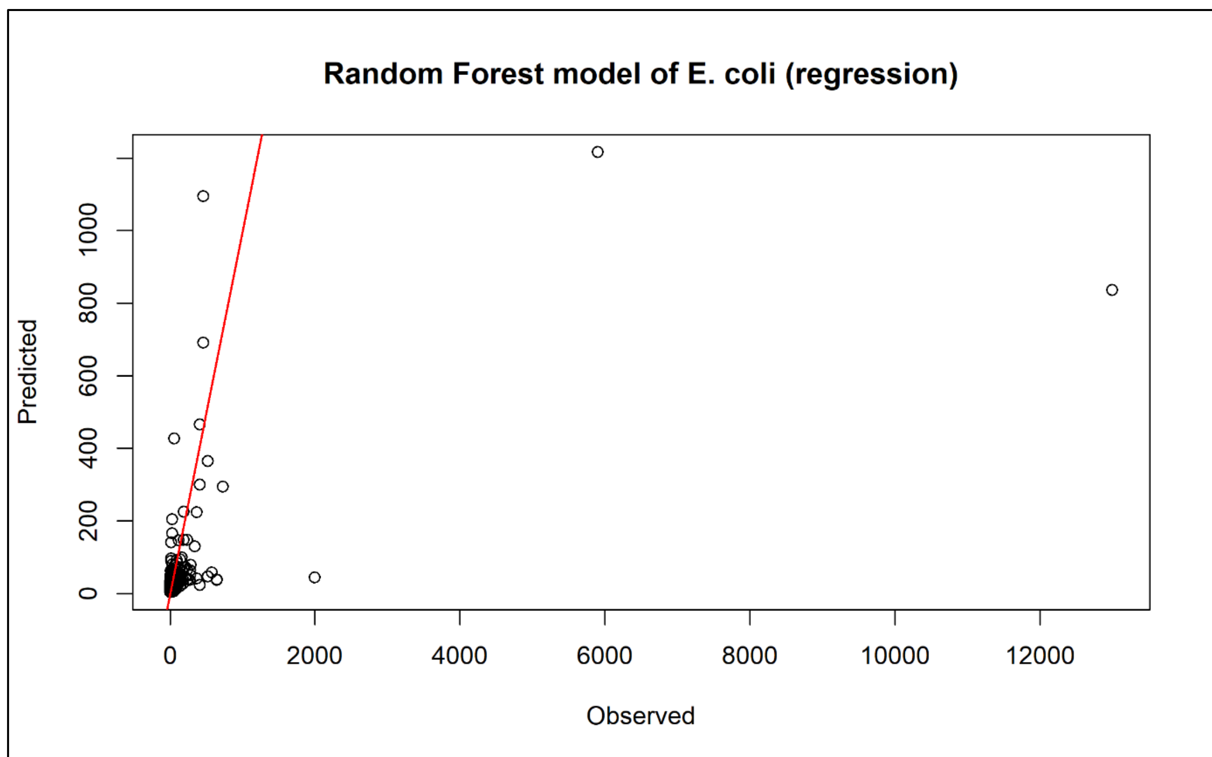


Figure 2.15: Observed and predicted *E. coli* organisms 100mL^{-1} for the validation of the random forest regression model. The red line indicates a 1:1 gradient, which would be a perfect prediction of the observed values.

2.3.2.3 Classification model

The optimised random forest model on the binary *E. coli* variable tried four variables at each split and had an OOB estimated error rate of 9.5%. Class error rates for the calibration dataset were 0.01 (under certification limit) and 0.56 (limit exceedance). Rainfall and turbidity were again the most important variables in the model, in terms of mean decrease in accuracy and Gini coefficient (Figure 2.16). While latitude was the third-most important variable in terms of predictive performance when considering the mean decrease in accuracy, it is important to note that it was not as important to the Gini index.

Rainfall and turbdity had similar modelled relationships with *E. coli*, as shown in their partial dependence plots (Figure 2.17); as rainfall and turbidity decreased, so too did the modelled probability of *E. coli* remaining under the certification limit, before plateuing at approximately 20mm and 30 NTU, respectively. The modelled relationship between *E. coli* and latitude was again more complex and highly variable.

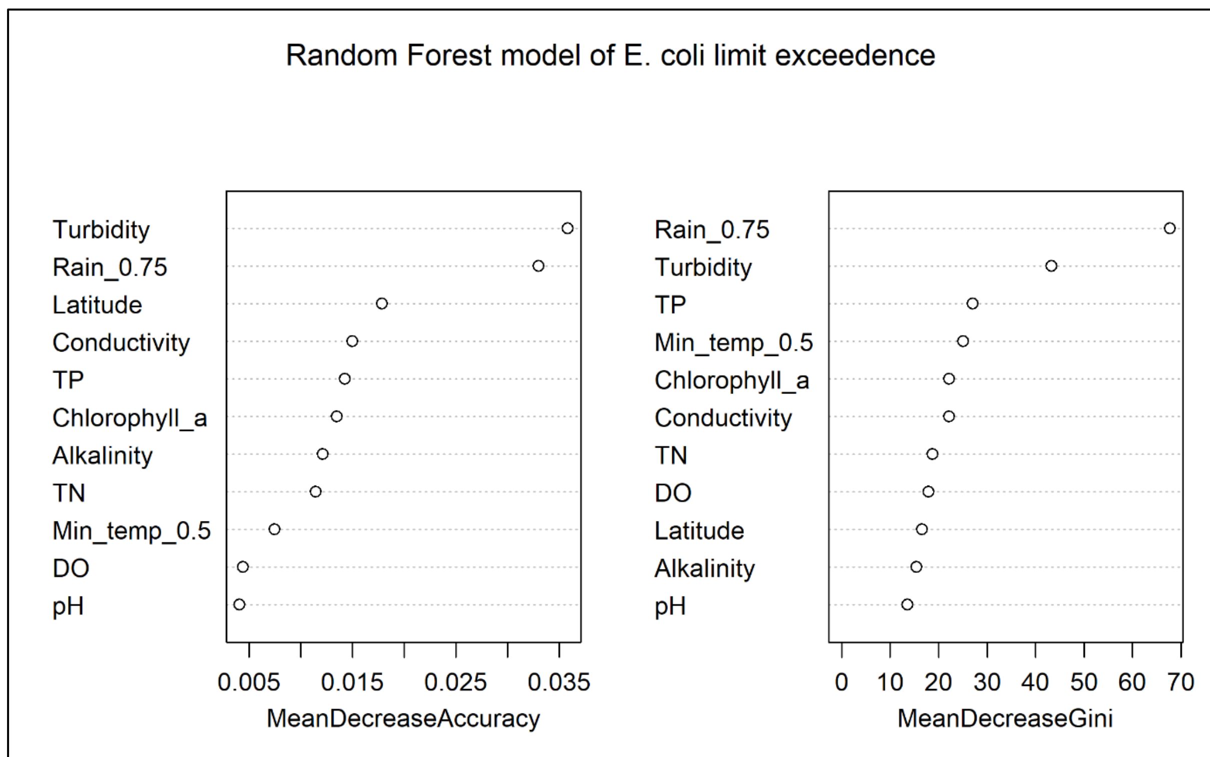


Figure 2.16: Variable importance plots for the classification random forest model of *E. coli* concentrations exceeding the Freshcare certification requirement of 100 cfu 100mL^{-1}.

The model performance evaluation on the withheld data resulted in the following values: accuracy = 0.90; specificity = 0.75; precision = 0.99; recall/sensitivity = 0.91; false positive rate = 0.25; false negative rate = 0.09. The area under the curve for the ROC graph (Figure 2.18) was 0.86.

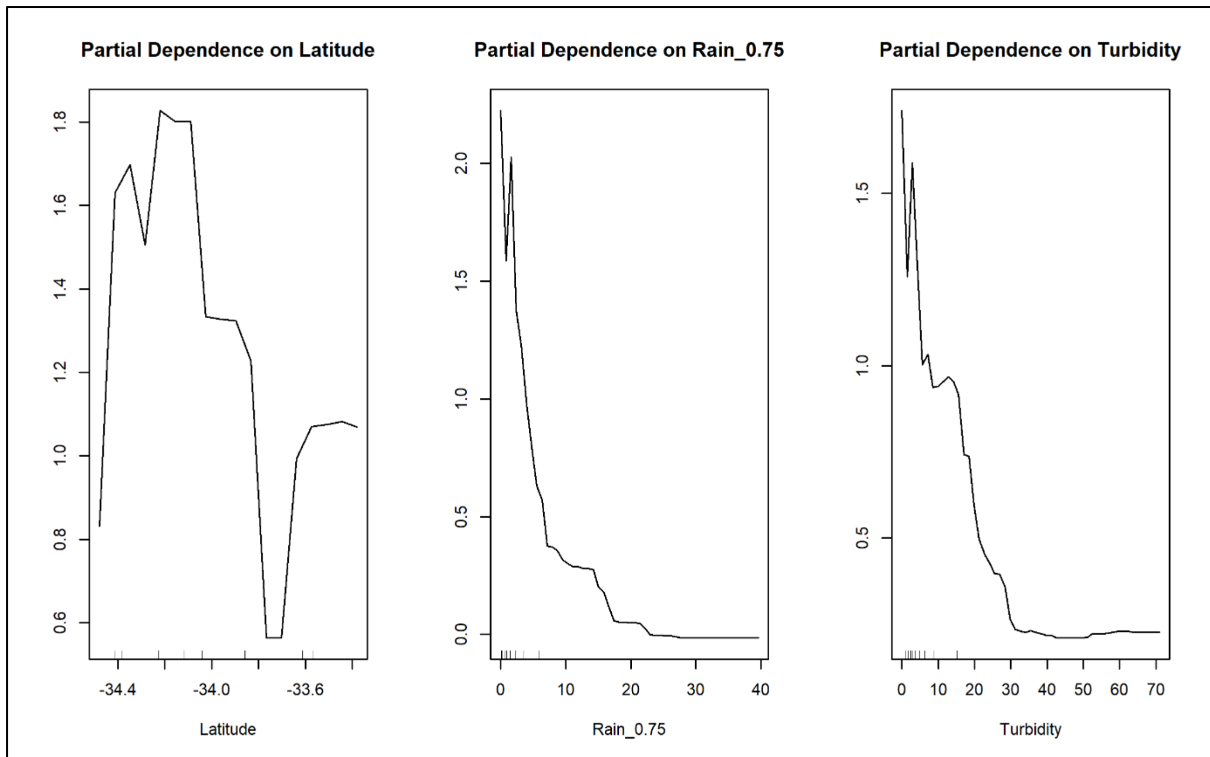


Figure 2.17: Partial dependence plots of the classification (binary E. coli) random forest model

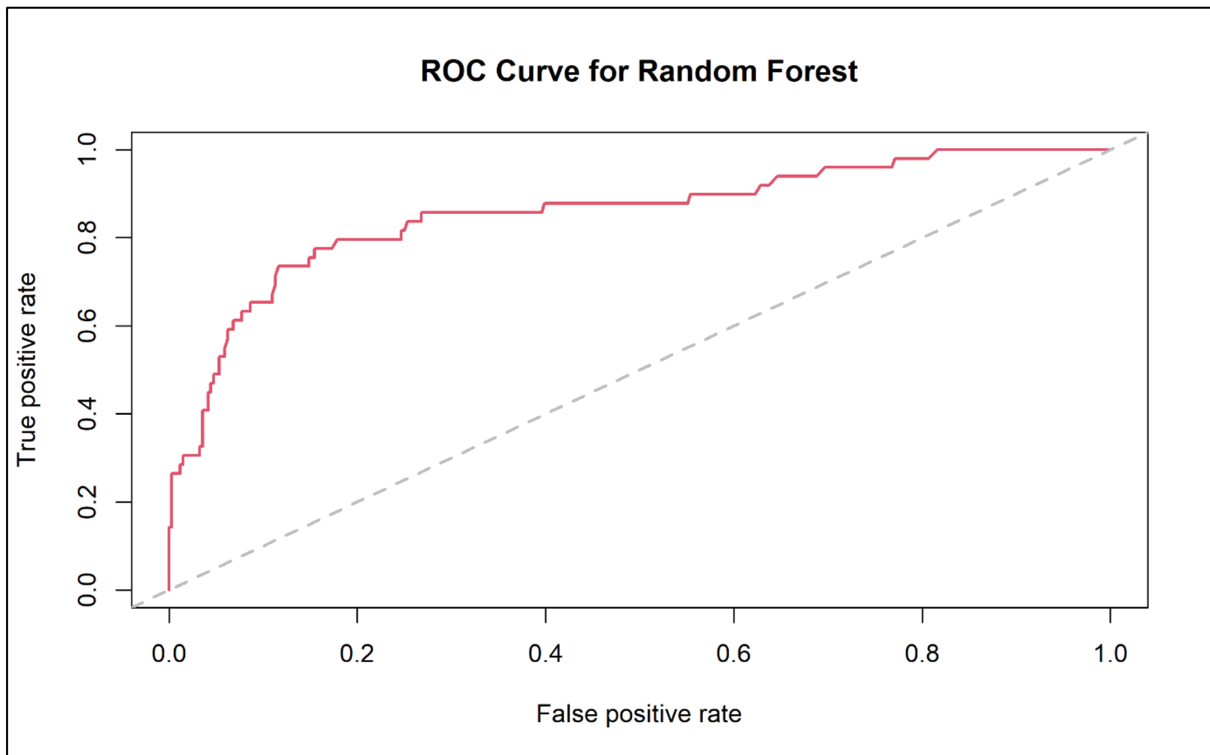


Figure 2.18: The receiver operator characteristic (ROC) graph for the random forest model on the binary E. coli variable (i.e., exceeding the Freshcare requirement). The red line indicates the actual performance of the model, with an area under the curve of 0.86. The grey dashed line indicates a 1:1 gradient of the true positive rate (TPR) and the false positive rate (FPR), or a completely random classification; perfect classification models would have a TPR of 1 and a FPR of 0, and the ROC would end in the top left of the graph.

2.4 Discussion

Escherichia coli (*E. coli*) is the microbial indicator used by the Australian fresh produce industry to determine irrigation water quality and safety. Growers measure *E. coli* in their water sources to ensure that they comply with certification requirements. However, infrequent sampling places uncertainty on the safe harvesting of fresh produce. Establishing strong correlations between *E. coli* and environmental and physicochemical parameters could allow for the prediction of microbial risk in real-time between laboratory verification (Herrig et al., 2015; Islam et al., 2018; Rock et al., 2016). The above statistical modelling shows that, despite evidence of strong individual correlations between water quality and environmental variables, there are still constraints around successfully predicting both *E. coli* concentration and certification limit exceedances.

2.4.1 Sediments

The purpose of incorporating sediments into the current study is because the sediments of streams and rivers may store high concentrations of faecal indicator bacteria and pathogens (Droppo et al., 2009) that can be mobilised during resuspension events (Bradshaw et al., 2016; Haller et al., 2009). This is relevant to fresh produce production because *E. coli* and the bacteriophage MS2 (a substitute organism for norovirus) have been shown to resuspend from irrigation water canals (Zhou et al., 2017), suggesting that pathogens may also resuspend. Further, studies have found the *E. coli* load in sediments can be 10- to 1000-fold higher in sediment than in the water above it (Benjamin et al., 2013), reiterating the importance of considering water-sediment interactions when evaluating the safety of irrigation water sources. By fitting the ATS trendline to the partially censored *E. coli* concentrations, the Hawkesbury study showed that there is a significant ($P < 0.01$) positive relationship between the concentrations in the water and in the sediment. This further emphasises that the sediment is both a source of and a sink for bacterial contamination in the water column. However, the persistence of FIB and genetic markers in sediment can vary within a watershed (Zimmer-Faust et al., 2017), and overall, pathogens are more frequently detected in water than in sediment (Bradshaw et al., 2016).

Published research has shown that including sediment features such as the concentration of *E. coli* in sediment and sediment particle size improves models for irrigation water sources (Tousi et al., 2021). The partial censoring of the sediment *E. coli* concentrations in the Hawkesbury field study prohibited their inclusion in the GAM. Additionally, due to the predominantly sandy soil type of the sites in this study, no significant effects of sediment particle size could be determined using the chosen statistical approaches.

2.4.2 Spatiotemporal variation and climate factors

Temporal and spatial variation was evident in the analysis of both datasets. Temporal variation was explored directly by examining season as specified by calendar dates in the Hawkesbury dataset, but the influence of season can also be seen through the importance of climate variables. Spatial variation was explored in several ways: the water-sediment interface as above, longitudinal and latitudinal gradients, and in comparing on-river to on-farm sampling sites.

For the Hawkesbury dataset, the Peto-Peto nonparametric tests concluded that there was no significant seasonal trend for *E. coli* concentrations in the water, but there was in the sediment; the tests showed that sediment *E. coli* concentrations were highest in autumn and summer, and lowest in spring and winter. This is an unexpected result because previous many studies have found significant seasonal trends for FIB in water (Carter et al., 1987; Cho et al., 2016; Nguyen et al., 2016; Paule-Mercado et al., 2016; Riedel et al., 2015; Sibanda et al., 2013). Further, the sediment was expected to be more resistant to seasonal fluctuations as they generally represent longer-term conditions (Nevers et al., 2020). The significant seasonal trend in the sediment bacterial loads may be a result of seasonally-dependent processes such as surface run-off (Cha et al., 2016), whereby the 'rainy' season delivers more bacterial contaminants to a water source due to increasing overland flow. With the suggestion of seasonality in the data, it is important to consider patterns in temperature, rainfall, and solar radiation exposure when developing the statistical models.

Due to multicollinearity (the GAM) and concurvity (the random forest models), solar exposure was not specifically modelled in this study. However, it was correlated to ambient temperature, which had a significantly positive effect on the concentration of *E. coli*. Similarly, discount factors of rainfall selected for model inclusion reveal more about the impact of these antecedent weather conditions. A discount factor (DF) close to 0.5 is equivalent to the current weather conditions; a DF of 0.99 is essentially representative of average conditions, similar to a moving average of rainfall (Wang et al., 2011). The temperature DF selected for the GAM and the random forest models was 0.5, suggesting that day-of temperature – maximum or minimum – had the greatest impact on *E. coli* concentration. For rainfall, a DF of 0.99 was selected for the GAM and 0.75 for the random forest models. This suggests that an accumulation of the preceding rainfall records had the greatest impact on *E. coli* concentration.

The relationships between *E. coli* and these meteorological parameters are variable and often complicated. The relationship between temperature and *E. coli* in water is complex due to the variability in inactivation rates between different water source types, meaning that the

importance of temperature appears to be site-specific (Blaustein et al., 2013). Studies have found that water and ambient temperature can be both positively correlated (Holvoet et al., 2014) and negatively correlated (Falardeau et al., 2017) with human pathogen occurrence.

Conversely, the correlation between rainfall and the microbial contamination of water is usually positive but modest and can be inconsistent among season, site, and microorganism (Wilkes et al., 2009). In one study, pathogen occurrence was most strongly correlated with higher levels of total precipitation over the five days before sampling, and specifically on the day just before sampling (Falardeau et al., 2017). Another study found that cumulative rainfall had a strong relationship to *E. coli*, more so than on the day of sampling (Wilkes et al., 2009). The models behind the function of the AgWater App include the total rainfall on the day of sampling and in the two preceding days as significant variables (Rock et al., 2016), with rainfall on the day prior to sampling being the most impactful. The importance of this lag could be related to travel distance from the sampling point to known pollution sources, which may be a useful factor to explore in future studies. Another study found that there was no correlation with rainfall at all (McEgan et al., 2013). These contrasting findings could be due to the geography of the catchment in which they were set, the number and size of pollution sources, or differing patterns of rainfall. Other studies have found that *E. coli* levels increased with windspeed (Benjamin et al., 2013), antecedent dry days (Paule-Mercado et al., 2016), and 3-days-sum solar radiation (Herrig et al., 2015).

2.4.2.1 Site-specific influences

The importance of longitude in the Hawkesbury dataset, which encompasses a west-to-east flowing section of river, and the importance of latitude in the Greater Sydney dataset, which encompasses a south-to-north flowing section of a river system, strongly suggests that there are site-specific factors influencing the microbial contamination of the water. This point is further emphasised by the complexity within the partial plots for the GAM and the random forest models.

Site was a substantially significant variable in predicting generic *E. coli* in an agricultural region of Arizona, but was not largely important in New York (Weller et al., 2020b). Site differences could be explained by adjacent land use or non-point pollution sources (Reitz et al., 2021). FIB has been predicted to increase with urban development, more so than for agricultural uses (Paule-Mercado et al., 2016). Within agricultural areas, the highest *E. coli* concentrations have been found in drainage ditches, followed by canals (Truchado et al., 2017); large water reservoirs have the lowest risk. Other studies have confirmed that the highest *E. coli* risk is associated with standing water (Benjamin et al., 2013), which relates to the comparison of on-farm and on-river sites. However, no significant differences were found

in the mean *E. coli* concentrations between the different source types in the Hawkesbury River region. It is important to note that the Greater Sydney dataset only included on-river sites, and therefore could not supply more statistical power to this analysis. Further investigation into the exact pollution sources for each of the sites could be warranted here. For example, one study found that *E. coli* correlates to cattle in summer and waterfowl in autumn (Hansen et al., 2020), also highlighting the variable importance of the animal hazards observed.

2.4.3 Physicochemical variation

Turbidity often reflects an increase in nutrients and surface run-off and therefore has commonly been associated with increased microbial contamination in water sources (Herrig et al., 2015; Paule-Mercado et al., 2016; Weller et al., 2020b). This was shown in both the GAM and the random forest models. For both statistical approaches, there was an increase in *E. coli* with increasing turbidity, followed by a plateau. However, this plateau occurred at substantially different points; 100 NTU for the GAM and 30 NTU for the random forest models. It is likely that this difference relates to the typical turbidity readings within each dataset; the Hawkesbury samples commonly had values above 50 NTU, while the Greater Sydney samples were mostly below 10 NTU.

Dissolved oxygen (DO) was significantly correlated with *E. coli* in the GAM but was not an important variable in the prediction of *E. coli* in the random forest models. *E. coli* concentration was highest at around 100% DO saturation, which is expected for an aerobic bacterium. Like turbidity, the Hawkesbury dataset had greater variation in DO compared to the Greater Sydney dataset, which might account for why the model found DO to be significant. Other studies have also found DO to be significant (Weller et al., 2020b, 2020a).

A contrasting observation can be made for conductivity; modelled *E. coli* dipped at approximately the median conductivity reading. The reason for this trend is unclear. *E. coli* can correlate negatively with electrical conductivity if diluted by rainfall (McEgan et al., 2013), but this was not evident in either dataset. The pairwise correlations performed prior to modelling did not suggest that multicollinearity between longitude and conductivity was problematic, but it is possibly still influencing the analysis. The Hawkesbury River becomes tidal and therefore saline in the northern reaches, creating outliers for the parameter of electrical conductivity. A fair amount of flexibility in the smooth terms for conductivity (edf=3.68) and longitude (edf=5.68) were required to fit the data. Parameter variability in a statistical model can cause this over-fitting because it allows the model to fit the noise in the data instead of the underlying pattern.

2.4.4 Model performance

Our ability to statistically model FIB, or more broadly any bacteria, is limited because of the highly context-specific persistence in the environment, driven by replication and die-off (Oliver et al., 2006). In turn, inadequate methods for accurately predicting *E. coli* occurrence in water continues to limit the identification of appropriate and effective fresh produce contamination prevention methods (Hansen et al., 2020).

The GAM was able to explain a reasonable amount of variation in the Hawkesbury data (62.9%), but the amount of flexibility allowed in the smooth terms (e.g., 5.68 estimated degrees of freedom for longitude) implies that the model was potentially overfitting. Random forest models can also be reliable tools for the analysis of water quality, as shown by the regression model explaining 57.5% of the training data variation. However, the performance of the random forest models when predicting onto the validation subset is more pertinent to evaluate. The regression model had an R^2 value of only 0.15 and failed to predict higher concentrations of *E. coli*. The classification random forest model of the binary term of the Freshcare certification limit exceedance performed better during the validation, making it more appropriate for application outside of this study. The superior performance of classification machine learning over regression machine learning when modelling FIB has been found previously (Tousi et al., 2021).

Importantly, the false negative rate (FNR) of the classification model validation was 0.09 and the false positive rate (FPR) was 0.25. This means that there is a 9% chance the model would incorrectly predict a water source as being acceptable for direct-contact irrigation when it exceeded the microbial limit, and a 25% chance that the model would warn against the use of the water source when it was within acceptable microbial limits. So, if used in practice, this model could give false confidence in water safety despite performing well from a statistical perspective.

It was important to perform a true validation of the predictive model (i.e., validation on data independent to the model's calibration) because this informs the utility of applying water quality models more broadly (Daggupati et al., 2015). Interestingly, it does not appear that the models behind the AgWater App were independently validated (Rock et al., 2016).

2.4.5 Limitations and further research

There is a smaller evidence-base for models predicting FIB than for other agricultural pollutants such as nutrients (Oliver et al., 2016a). The dynamic nature of *E. coli* concentrations makes them difficult to model (Harmel et al., 2016; Oliver et al., 2016b). Studies have had some success with using artificial neural networks (ANNs) to predict FIB

concentrations along beaches (Sulaiman et al., 2019; Zhang et al., 2018), but a comparison of modelling approaches found that the performance of ANNs can vary along stretches of the same river (Mälzer et al., 2016). In another study, hierarchical Bayesian regression models using antecedent meteorological conditions for predicting FIB in urban waterways was able to classify ‘safe’ and ‘unsafe’ levels with good accuracy (Farnham and Lall, 2015). Bayesian models have also been applied to the prediction of environmental *E. coli* blooms (Bertone et al., 2019). Alternative statistical approaches such as these should be considered for irrigation water sources, with the suitability, quantity, and quality of the data being used as an important consideration.

The Hawkesbury study was based on a relatively small dataset. Only 96 observations were made during 12 months of study, though the number of water quality and environmental parameters complimentary to the *E. coli* measured is reasonable. An assumption of this study is that the sampling period is representative of “normal” conditions, as it encompasses only one iteration of each season. Future research could focus on collating larger datasets to improve the statistical power, either through increasing the time period (i.e., the Greater Sydney dataset) or increasing the frequency of sampling over the 12 months (Herrig et al., 2015). Further, if the connection between microbial contamination in the water and the sediment is to be explored in more depth, it would be beneficial to study an increased range of different soil types to better elucidate the importance of the sand, silt, and clay percentages, as has been proven in lakes (Chandran et al., 2011) and estuaries (Craig et al., 2004).

2.5 Conclusion

Aside from the two tributaries of the Grose and Nepean Rivers, all the Hawkesbury water sampling sites had mean *E. coli* concentrations close to or above the Freshcare certification requirement of 100 cfu per 100mL. This implies that irrigation water sources in the area would place the safety and quality of fresh produce at risk. Further, the significant relationship between *E. coli* in water and sediment should reinforce the importance of considering microbial loads in sediment when managing the quality of irrigation water sources. The GAM and random forest models in this study align with the principles behind the AgWater App in that a combination of physicochemical and environmental parameters are necessary when developing robust models of irrigation water risk (Rock et al., 2016). Overall, the sampling sites were representative of a typical mixed-use peri-urban area, meaning there are wider implications for similar peri-urban irrigation water sources around Australia.

3 Bacterial community composition and food safety risks in irrigation water sources

3.1 Introduction

As next-generation sequencing (NGS) has become more economical and accessible, freshwater microbial communities have been characterised in depth (Costa et al., 2015; Iliev et al., 2017; Llirós et al., 2014). Freshwater environments have been found to be exceptionally biodiverse in a range of studied water sources, including sources used for agricultural irrigation (Abia et al., 2018; Allard et al., 2019; Rusiñol et al., 2020). Only a subset of taxa within these diverse communities will be of concern to the fresh produce industry, due to the human health implications of food becoming contaminated with pathogens during irrigation (Oliveira et al., 2012). However, accurately isolating, culturing, and identifying even a subset of taxa in irrigation water sources can be expensive and time-consuming. The recent advancement of molecular tools offers an alternative approach to traditional, culture-based methods when classifying the microbial risk of water sources (Cao et al., 2017).

NGS has been used to study the taxonomic and functional profiles of a variety of freshwater environments (Nevers et al., 2020; Roberto et al., 2018; Shang et al., 2022; Shilei et al., 2020). Previous studies of surface water sources in agricultural areas have used NGS to show that bacterial community composition and function are closely related to land use (Abia et al., 2018; Chen et al., 2018; Staley et al., 2014) and is more variable in smaller streams or channels (Chen et al., 2018). But relatively few studies have classified the risk of irrigation water as it relates to fresh produce food safety, as reviewed in Jagadeesan et al. (2019), despite the success of the available techniques in identifying a wide range of waterborne pathogens.

Rusiñol et al. (2020) studied human pathogens from fresh produce irrigation water sources in Spain using NGS for bacteria, viruses, and protozoa. The authors found that river sources were contaminated with *Pseudomonas*, *Flavobacterium*, *Bacillus*, *Enterobacter,* and *Serratia*, regardless of the season, and detected *Naegleria clarki* in summer and Norovirus in autumn. This study demonstrates that NGS is becoming a useful tool to detect human pathogens from environmental genetic material and reaffirms that sources of pre-harvest water are a critical human-health concern.

The following chapter applies 16S rRNA sequencing to water and sediment samples to identify potential food safety concerns that are present in fresh produce irrigation water

sources. This chapter describes the bacterial communities present in the studied samples by generating taxonomic and functional profiles as assigned by the SILVA genomic database (Quast et al., 2013) and inferred by the KEGG genomic database (Qiu, 2013), respectively. This chapter also introduces the concept of an irrigation pathogen potential index of bacterial genera identified in the sequences as an indicator of fresh produce food safety risk.

3.2 Methods and materials

3.2.1 Sampling

The sampling design for this chapter was as described in Chapter 2 (page 16). Water and sediment samples were collected monthly from the upper Hawkesbury River area in Greater Sydney, New South Wales, from September 2018 to August 2019. In total, 190 samples for metagenomic analysis (96 water, 94 sediment) were collected from eight sites (five river, three on-farm storages) over a 12-month period.

3.2.2 Sample preparation and DNA extraction

Collected samples were first processed for *Escherichia coli* enumeration, as described in Chapter 2 (page 18). DNA was then extracted directly from the sediment samples using the DNeasy PowerSoil Pro Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Due to the high suspended clay content of many of the sample sites (Table 2.1), the water samples required a multi-step concentration process: flocculation, centrifugation, and filtration. The flocculant helped to separate any suspended sediment from the water column, did not have measurable antimicrobial effects at the concentrations used, and did not impede DNA extraction processes further down the workflow. This method was adapted from Karanis and Kimura (2002).

Aluminium potassium sulphate ($KAl(SO_4)_2$) was dissolved in autoclaved MiliQ water and used as a flocculant. Each 1-litre water sample bottle received a minimum of 100 mg equivalent of flocculant; samples that were very turbid received up to 200 mg equivalent to ensure a robust floc. After adding the flocculant, the bottles were inverted several times to mix. The samples were left undisturbed at 4°C for 12-24 h to allow the flocculant to take effect.

Flocculated samples were then centrifuged in approximately 1-litre amounts in O-ring bottles (Nalgene, Rochester, NY, USA). The samples were centrifuged for 20 minutes at 10,000 *g* using a Sorvall RC6 centrifuge (Thermo Fisher, Waltham, MA, USA). Following centrifugation, the supernatant was carefully decanted into sterilised filter funnels and the pellet was retained. The pellets were combined to create a single composite pellet for each

site. The pellets were then transferred to sterile 50mL conical centrifuge tubes (Falcon-Corning, Corning, NY, USA) and centrifuged at 10,000 *g* for 5 minutes to ensure concentration of the bacteria. DNA was extracted from the pelleted samples using DNeasy PowerSoil Pro kits (Qiagen, Hilden, Germany).

The remaining supernatant was filtered in a vacuum manifold (Merck Milipore, Burlington, MA, USA) powered by an electric pump. The glass collection funnels and filter heads were sterilised between the processing of each site using a bleach solution that was then sufficiently neutralised with sodium thiosulfate. Nylon filters of 0.45µm pore size (Merck Milipore, Burlington, MA, USA) were used until they reached capacity (i.e., no flow-through for at least 5 minutes), or until all the water was filtered. The number of filter papers for each site varied depending on the turbidity that remained after flocculation. Each filter paper was then eluted in 1mL of sterile 1X PBS (pH 7.4) prior to DNA extraction. The eluates for each sample were combined in sterile tubes (Falcon-Corning, Corning, NY, USA) and centrifuged at 15,000 *g* for 1 minute. The supernatant was discarded, and the pellet was resuspended in 500µL of sterile 1X PBS (pH 7.4) using a micropipette. Finally, 500µL of eluate from each concentrated sample was then processed using a PowerSoil DNA extraction kit (Qiagen, Hilden, Germany).

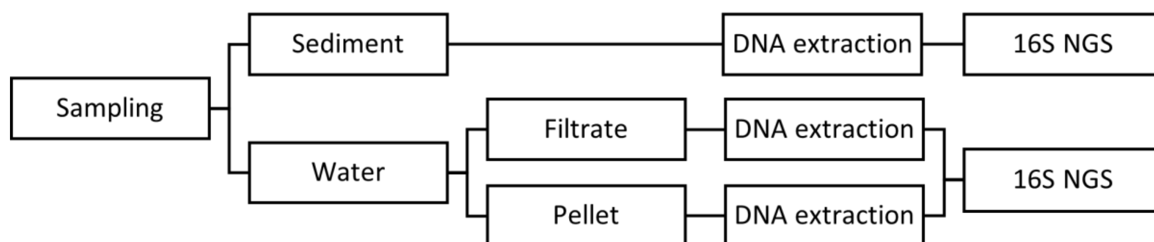


Figure 3.1: Diagram of sample preparation for DNA extraction and 16S sequencing. DNA was extracted from the sediment samples directly. The water samples underwent concentration (i.e., flocculation, centrifugation, and filtration) prior to DNA extraction.

The purity and estimated concentration of the DNA extracts were measured using a NanoDrop Microvolume Spectrophotometer (Thermo Fisher, Waltham, MA, USA) to ensure the DNA was sufficient for sequencing. Approximately 20ng of DNA from each sediment sample and a combined total of 20ng of pellet and filter paper DNA for each water sample were processed for metagenomic profiling, as shown in Figure 3.1. All DNA extracts were stored at -80°C until processed.

3.2.3 Next-generation sequencing (NGS) and data processing

NGS was performed for the sediment and water DNA (Figure 3.1) by the Ramaciotti Centre for Genomics (ISO/IEC 17025) at The University of New South Wales (Sydney, NSW,

Australia). The 27F/519R primer set (Lane, 1991) was used to target the V1-V3 hypervariable region of the bacterial 16S rRNA gene. The amplicons were sequenced on an Illumina MiSeq platform (Illumina, San Diego, CA, USA).

The raw, demultiplexed fastq files were quality filtered and paired using the 'dada2' package (Callahan et al., 2016a) in R (R Core Team, 2021). Sequences were truncated at 290 bp (forward) and 270 bp (reverse), with maximum allowed errors limited to 3 (forward) and 4 (reverse). Reads were filtered to a minimum quality score of 5. Error learning was performed separately on the amplicons from sediment and water samples. Amplicon sequence variant error correction was performed with the DADA2 pipeline in the 'dada2' package (Callahan et al., 2016a). Sequencing errors were resolved with the in-built denoising algorithm to generate true amplicon sequence variants (ASVs), which are more precise and reproducible than operational taxonomic units (OTUs) (Callahan et al., 2017). Following sample inference, the filtered sequences were merged, which also removed any sequences that did not have an exact overlap. Chimeric sequences were then removed using the in-built *de novo* method within the DADA2 pipeline, which identifies chimeras by consensus across samples (Callahan et al., 2016b).

3.2.4 Taxonomic and functional profiling

The ASVs were aligned to taxonomies in the SILVA database (version 138) (formatted by McLaren and Callahan, 2021). The resulting taxa were filtered to remove matches for the Eukaryota kingdom and mitochondrial RNA as classified under the Rickettsiales order in the SILVA database (Quast et al., 2013).

The 10,000 most abundant ASVs were used for functional profile prediction with Piphillin (Iwai et al., 2016), as has been previously shown to be an appropriate application for ASVs (Narayan et al., 2020). Piphillin uses the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (release 88) (Qiu, 2013) to infer metagenomic content from 16S rRNA sequences. The predicted metagenomic function of the sampled microbial communities was compared to KEGG with a 99%-identity requirement. Alignments with 'Global overview maps' in the 'Metabolism' tier were removed prior to plotting due to the substantial number of hits categorised in this tier and the lack of specific detail provided by the grouping.

3.2.5 Irrigation pathogen potential index (IPPI)

An irrigation pathogen potential index (IPPI) was developed from the annotated taxonomies. This is similar to the designation of sewage-associated taxa (SAT) as described in Nshimyimana et al. (2017). Bacterial genera of concern to the Australian fresh produce industry that are also known to transfer onto crops through irrigation water were selected.

The fifteen genera were *Aeromonas, Bacillus, Campylobacter, Clostridium, Enterobacter, Enterococcus, Escherichia, Klebsiella, Listeria, Pseudomonas, Salmonella, Shigella, Staphylococcus, Vibrio,* and *Yersinia*. The cumulative relative abundance of these genera within each sample was used to quantify the index. Additionally, the data was explored to determine the presence of potentially pathogenic species.

3.2.6 Data analysis

Data manipulation was performed with the 'dplyr' package in R (Wickham et al., 2022) to calculate relative abundances at relevant assigned taxonomic ranks and to produce summary statistics. All results were plotted with the 'ggplot2' package in R (Wickham, 2016). Pairwise comparisons of the irrigation pathogen potential index between the sample groups (i.e., water source, sample matrix, season) were evaluated using Wilcoxon rank sum tests (Wilcoxon, 1945).

The sampling program was designed to ensure that replicate sites were sampled for each water source (i.e., on-farm storage and on-river) to capture conditions that might be considered typical for that source type, as opposed to emphasising variation between individual sites. Applying metagenomic techniques at a site-specific level would be an advantageous research objective, but this study was conceived with the aim of supplying the fresh produce industry with more general information; blanket advice cannot be extracted from a study on site-by-site variation.

3.3 Results

3.3.1 Metagenomic sequence generation

A total of 10,146,372 sequence reads and 67,390 ASVs were obtained from 190 samples after processing and quality filtering through the DADA2 pipeline (Table 3.1). In general, samples with a larger number of sequence reads also recorded more unique ASVs, with the notable exception being water which had about a third of the number of unique ASVs as the sediment, but almost double the hits.

All ASVs were identified to the family level (Table 3.2) and a majority (66.1%) were identified to the genus level. However, only a very small percentage (0.61%) were able to be identified to the species level.

Approximately half (52.11%) of all sequence reads were attributed to the Proteobacteria phylum (Table 3.3), with the predominant class being Gammaproteobacteria in respect to both relative abundance (38.14%) and number of member genera (279). Other major phyla

included Bacteroidota (17.03%) and Actinobacteriota (13.55%). Cyanobacteria contributed a modest relative abundance (3.30%) but represented 100 member genera.

Table 3.1: The distribution of the number of samples, amplicon sequence variants (ASVs), and sequence reads from samples (n = 190) collected between September 2018 and August 2019.

	Samples	**ASVs**	**Reads**
Summer (Dec-Feb)	47	32,920	3,741,468
Autumn (Mar-May)	48	25,264	2,680,511
Winter (June-Aug)	48	21,319	2,224,053
Spring (Sept-Nov)	47	15,782	1,500,340
Water	96	18,672	6,578,140
Sediment	94	52,321	3,568,232
River (5 sites)	120	43,694	6,332,196
Storage (3 sites)	70	30,940	3,814,176
Total	**190**	**67,390**	**10,146,372**

Table 3.2: Number of amplicon sequence variants (ASVs) successfully mapped to taxonomic ranks in the SILVA (v138) database, including percent of total ASVs.

Kingdom	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
67,390 (100%)	67,390 (100%)	67,390 (100%)	67,390 (100%)	67,390 (100%)	44,545 (66.1%)	411 (0.61%)

The version of the SILVA database used in this study (version 138) (McLaren and Callahan, 2021b) had undergone a major upgrade prior to use, including reclassifying all Betaproteobacteria as Gammaproteobacteria, which limits the amount of information that can be gathered from the class level in this study.

Table 3.3: Number of 16S rRNA sequence reads, amplicon sequence variants (ASVs), relative abundance (%) of total sequence reads, and number of member genera identified in taxonomic groups across all samples (n = 190) with the SILVA (version 138) database.

Taxonomic group	**# reads**	**ASVs**	**Relative abundance (%)**	**Number of member genera**
Proteobacteria (phylum)	5,286,989	21,559	52.11	
Alphaproteobacteria	*1,417,563*		*13.97*	*186*
Gammaproteobacteria	*3,869,351*		*38.14*	*279*
Bacteroidota (phylum)	1,728,168	13,293	17.03	
Bacteroidia	*1,679,306*		*16.55*	*139*
Actinobacteriota (phylum)	1,374,596	4,023	13.55	
Actinobacteria	*1,300,096*		*12.81*	*112*
Cyanobacteria (phylum)	335,143	1,020	3.30	
Cyanobacteriia	*333,643*		*3.29*	*100*
Chloroflexi (phylum)	309,163	4,630	3.05	
Anaerolineae	*220,774*		*2.18*	*16*
Chloroflexia	*28,609*		*0.28*	*9*
Firmicutes (phylum)	141,012	2,190	1.39	
Clostridia	*67,932*		*0.67*	*69*
Bacilli	*66,654*		*0.66*	*41*

3.3.2 Taxonomic profiles

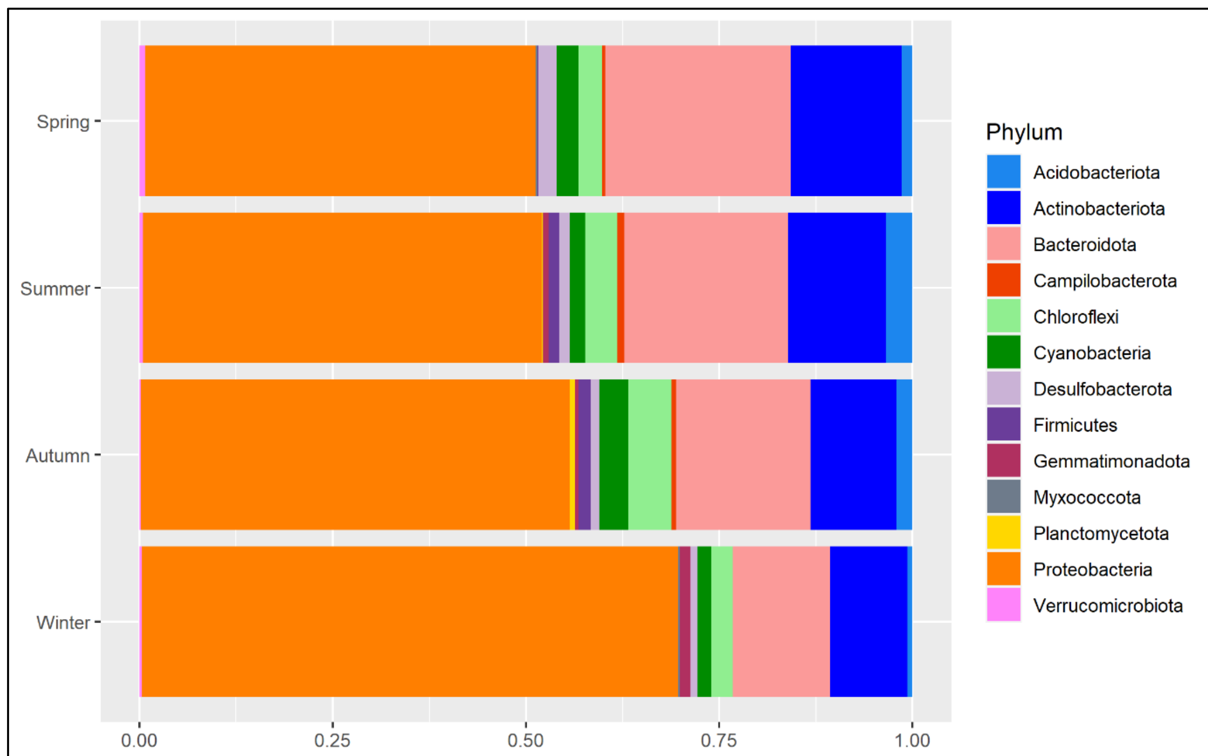


Figure 3.2: Relative abundance of phyla (>5%) in the sediment and water samples (n = 190) across the four seasons from September 2018 to August 2019 season (Spring: September-November; summer: December-February; autumn: March-May; winter: June-August)

The Proteobacteria phylum was the most abundant across all four sampling seasons (Figure 3.2). The relative abundance of Proteobacteria was slightly higher in autumn than in summer and spring, and notably higher in winter. The Bacteroidota phylum followed the inverse trend, being least abundant in winter and more prominent in summer and spring. The Actinobacteriota, Cyanobacteria and Chloroflexii phyla varied slightly throughout the seasons. Firmicutes, while not a major phylum, were more abundant in summer and autumn. Other phyla that contributed greater than 5% of the relative abundance included Acidobacteriota, Campilobacterota, Desulfobacterota, Gemmatimonadota, Myxococcota, Planctomycetota, and Verrucomicrobiota.

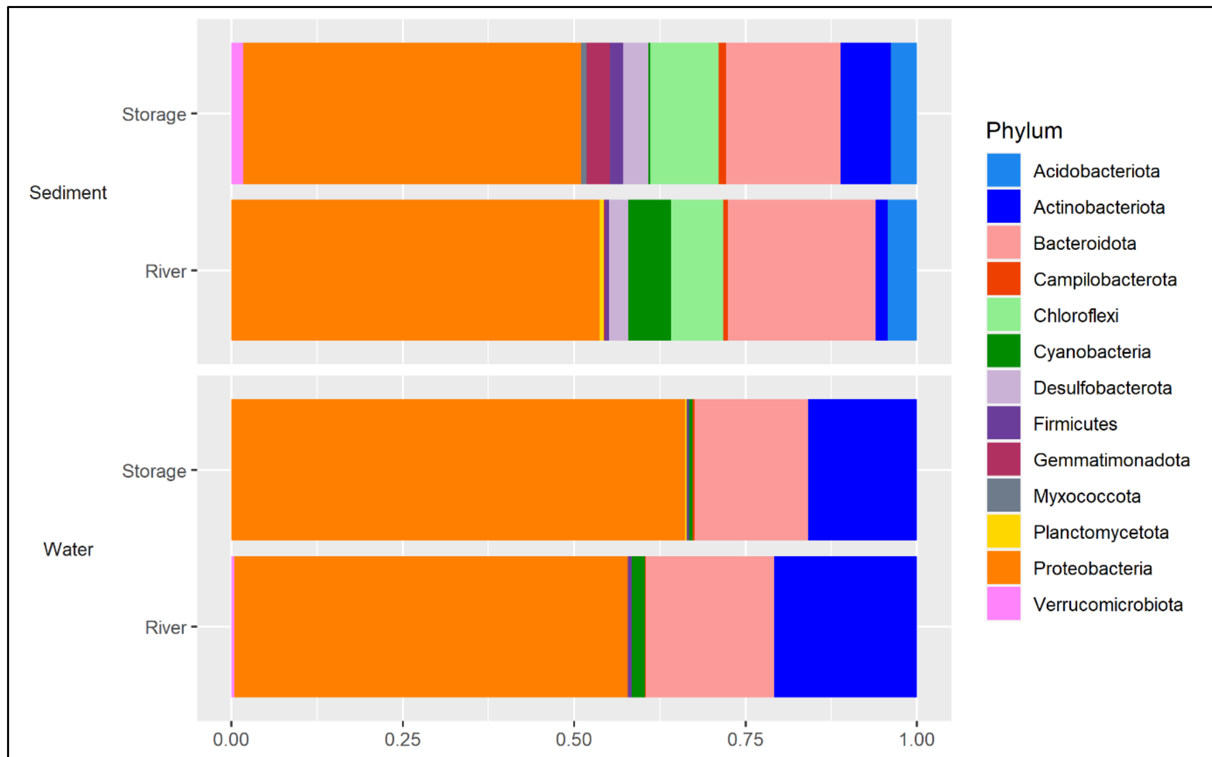


Figure 3.3. Relative abundance of phyla (>5%) in the sediment (n = 94) and water (n = 96) samples from on-farm storages (n = 3) and river sites (n = 5).

The Proteobacteria phylum was also the most abundant across both on-farm storage and river sites, and for both sediment and water sample matrices (Figure 3.3). Chloroflexi were notably more abundant in sediment, while Actinobacteria were more abundant in water. Cyanobacteria were more abundant at river sites, and most noticeably in the sediment. Overall, there appeared to be greater diversity at the phylum level in the sediment samples compared to the water samples.

Among the twelve most abundant orders within the Proteobacteria phylum (Figure 3.4), Burkholderiales were more abundant than any other order, contributing up to around 30% of relative abundance at this taxonomic rank. Rhizobiales were well-represented in the sediment samples, whereas taxa in the SAR11 clade were almost exclusively detected in the water samples. Across the four seasons, there was no substantial pattern except for a decrease in Burkholderiales and increases in several other orders during autumn.

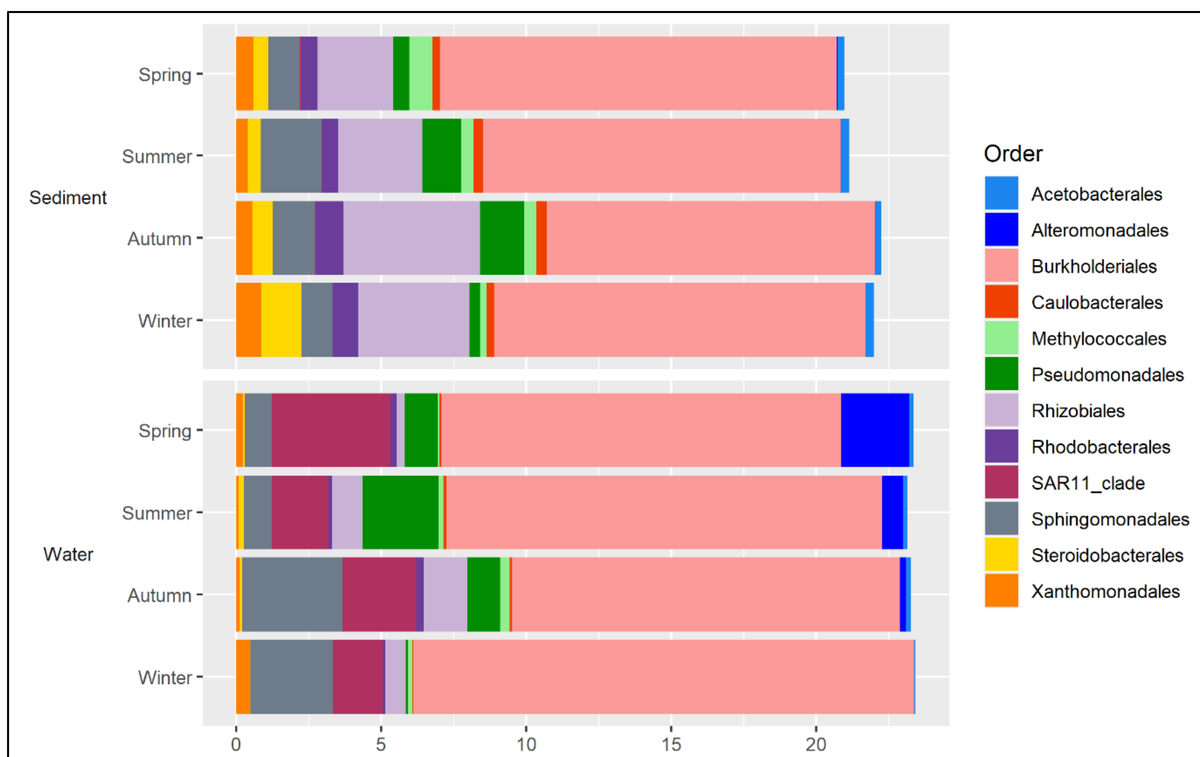


Figure 3.4: Relative abundance (%) of the top twelve orders within the Proteobacteria phylum in the water and sediment samples across the four seasons from spring, 2018, to winter, 2019.

The order Enterobacterales is not represented in Figure 3.4 because it contributed to only a small proportion of the overall bacterial community in this study (<0.1%). Figure 3.5 shows the variability in Enterobacterales throughout the study, both in overall representation and in diversity of member genera. The relative abundance was generally greater in the on-farm storages compared to the on-river sites. The relative abundance was also greater in the sediment than the water, but important differences can be seen at the genus level. *Escherichia/Shigella* (the SILVA taxonomy does not differentiate between these genera) was overwhelmingly dominant in the sediment from on-farm storage sites in winter. For example, at Yarramundi Lagoon in August 2019 (winter), 9.59% of sequence reads from the sediment sample were identified as belonging to the member species *Escherichia/Shigella coli* (data too insubstantial to be represented in Figure 3.5).

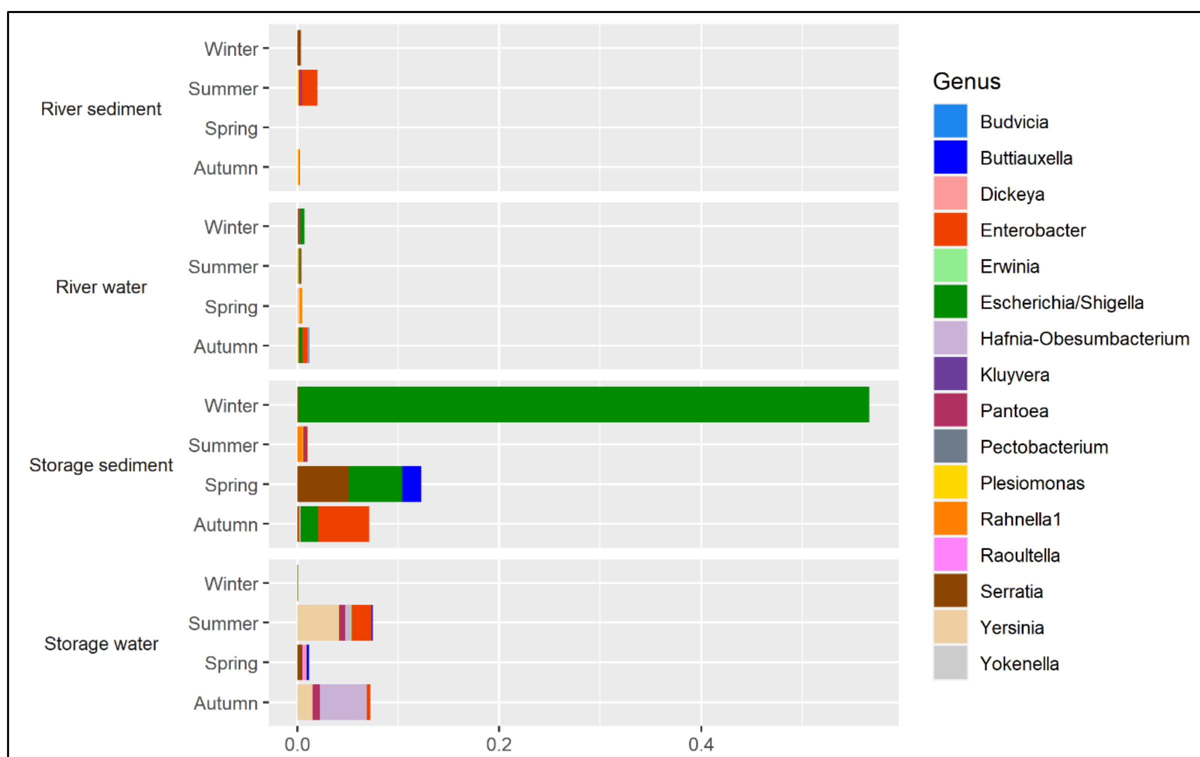


Figure 3.5: Relative abundance (%) of genera within the Enterobacterales order across the four seasons in the sediment and water samples from on-farm storages and river sites

3.3.3 Irrigation pathogen potential index (IPPI)

From a total of 44,545 ASVs identified to the genus level of taxonomic rank, 684 ASVs were identified as belonging to potentially pathogenic genera. Three genera of interest that were not identified in any of the samples were *Klebsiella, Salmonella,* and *Vibrio*. The genera of interest that were identified in at least one sample were *Aeromonas, Bacillus, Campylobacter, Clostridium, Enterobacter, Enterococcus, Escherichia/Shigella, Listeria, Pseudomonas, Staphylococcus,* and *Yersinia* (Table 3.4, Table 3.5). The combined relative abundance of these genera within each sample was used to form the irrigation pathogen potential index, as reported in Table 3.4 and Table 3.5, and assessed further in Figure 3.6 and Figure 3.7.

Overall, the mean relative abundances (%) of each of the potentially pathogenic genera were very low. *Pseudomonas* was the most abundant genus regardless of sample source and matrix (Table 3.4) or season (Table 3.5), being particularly abundant in on-farm storage water and in summer. *Aeromonas* and *Bacillus* were also reasonably abundant in the samples, but *Campylobacter* and *Listeria* were the least abundant.

Table 3.4: Mean relative abundances (%) of sequence reads attributed to each of the potentially pathogenic genera, separated by sample source and matrix (ND = not detected).

Genus	**River sediment**	**River water**	**Storage sediment**	**Storage water**
Aeromonas	0.21	0.35	0.12	0.41
Bacillus	0.07	0.03	0.37	0.02
Campylobacter	ND	<0.01	ND	ND
Clostridium	<0.01	ND	0.01	ND
Enterobacter	<0.01	<0.01	0.02	0.03
Enterococcus	ND	<0.01	<0.01	ND
Escherichia/ Shigella	ND	<0.01	0.87	ND
Listeria	<0.01	ND	ND	ND
Pseudomonas	1.61	1.54	1.58	4.14
Staphylococcus	<0.01	<0.01	0.16	ND
Yersinia	ND	<0.01	ND	0.08
Mean IPPI	**1.89**	**1.93**	**3.13**	**4.68**

Table 3.5: Mean relative abundances (%) of sequence reads attributed to each of the potentially pathogenic genera, separated by season (ND = not detected)

Genus	**Spring**	**Summer**	**Autumn**	**Winter**
Aeromonas	0.47	0.44	0.16	0.05
Bacillus	<0.01	0.20	0.16	0.04
Campylobacter	ND	ND	ND	<0.01
Clostridium	ND	<0.01	<0.01	<0.01
Enterobacter	ND	0.02	0.02	ND
Enterococcus	ND	<0.01	<0.01	<0.01
Escherichia/ Shigella	0.04	<0.01	0.01	0.57
Listeria	<0.01	ND	ND	ND
Pseudomonas	1.51	3.97	2.21	0.58
Staphylococcus	0.12	<0.01	ND	<0.01
Yersinia	<0.01	0.05	0.01	ND
Mean IPPI	**2.14**	**4.69**	**2.57**	**1.24**

Statistically significant trends ($P < 0.05$) were successfully obtained from the data. Boxplots of the log10-transformed irrigation potential pathogen index are shown in Figure 3.6 and Figure 3.7. Pairwise comparisons using Wilcoxon rank sum tests revealed that the IPPI was significantly higher in on-farm storage sediment than in river sediment ($P < 0.01$), was lowest in river water, and was the most variable in storage water. Additionally, samples collected in winter had significantly lower pathogen potential when compared to samples collected in other seasons ($P < 0.05$).

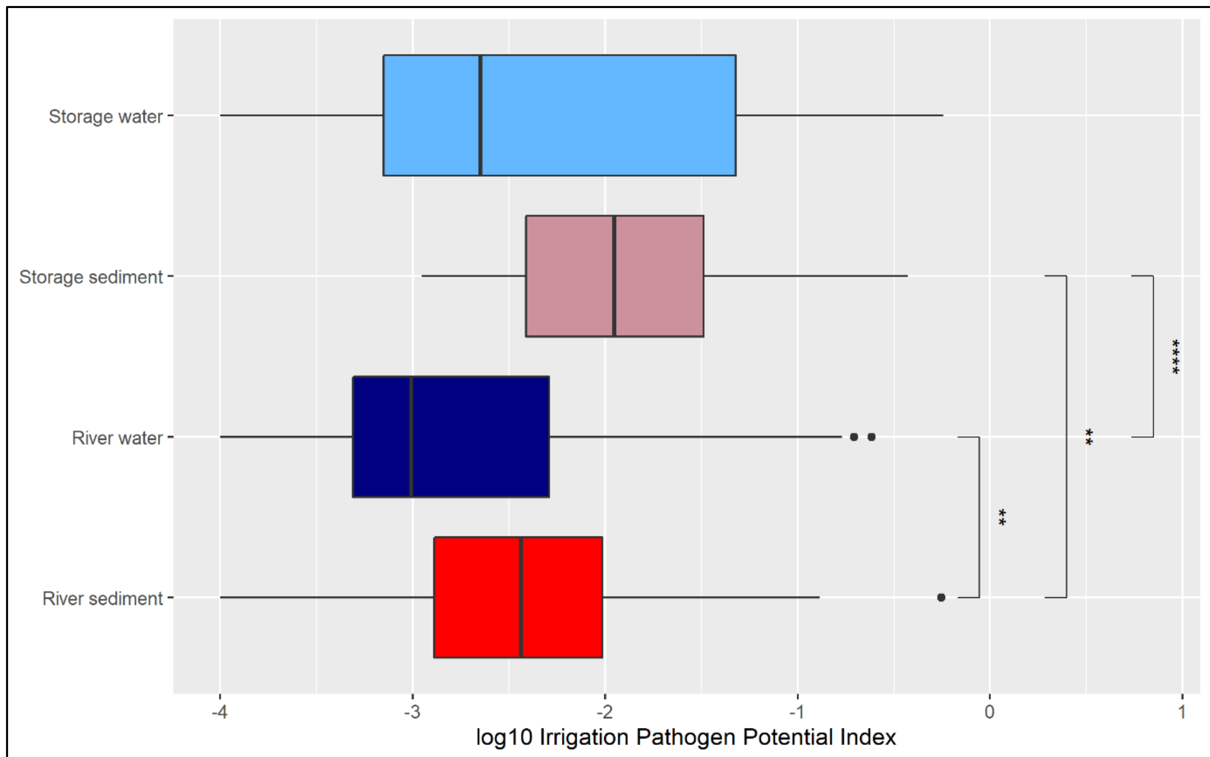


Figure 3.6: Boxplots of the log10-transformed irrigation potential pathogen index, separated by sample source and matrix, with significant comparisons shown. (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05)

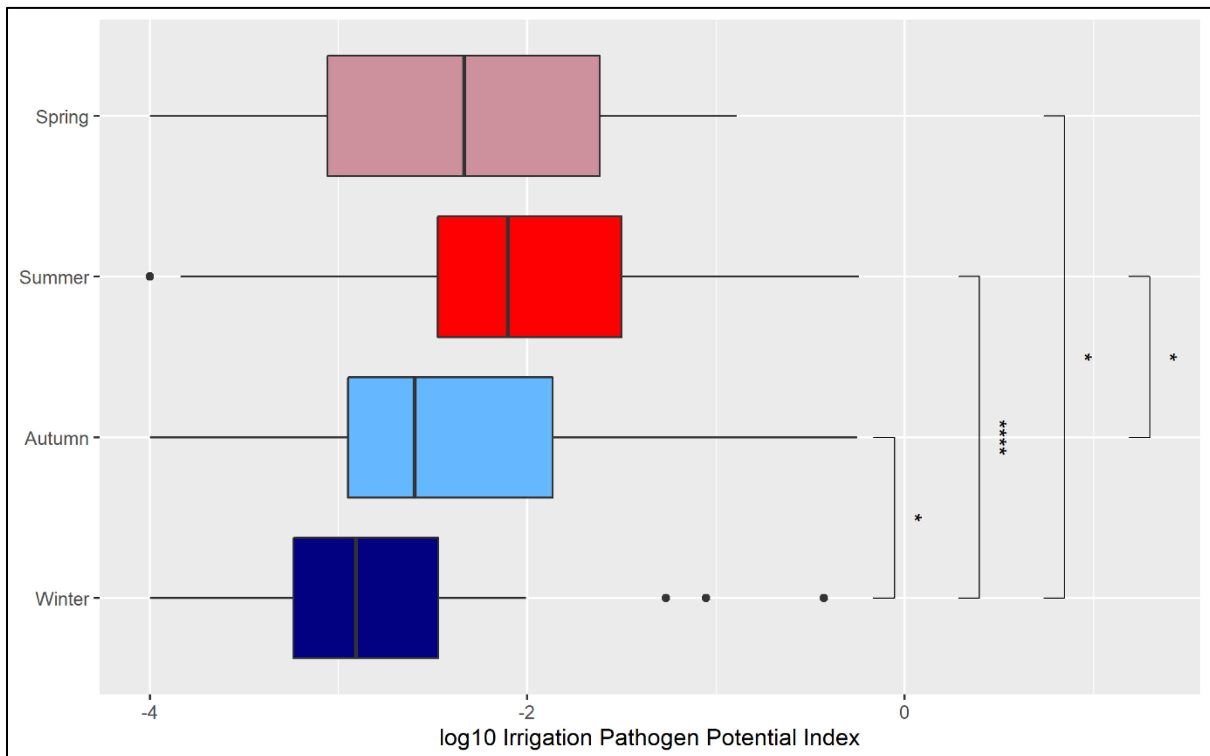


Figure 3.7: Boxplots of the log10-transformed potential pathogen proportions, separated by sampling season, with significant comparisons shown. (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05)

As previously stated, only a very small percentage of all the ASVs (0.61%) were able to be identified to the species level, including 60 ASVs within the irrigation pathogen potential

subset. Detected in a single sample each were *Bacillus cereus*, *Campylobacter coli*, *Listeria monocytogenes*, and *Staphylococcus aureus*. *Enterococcus faecium* was detected in two samples, and *Escherichia/Shigella coli* was detected in eight samples.

3.3.4 Functional profiles

The 10,000 most abundant ASVs were submitted to the Piphillin server for functional prediction, of which 481 ASVs had greater than 99%-identity to 16S genome sequences in the KEGG database. These ASVs inferred 191 genomes, 326 KEGG pathways, and 7080 KEGG features. Overall, greater functional prediction could be made for the water samples compared to the sediment samples (Figure 3.8). Most of the functional predictions (60.5%) related to metabolic processes, followed by genetic (13.1%) and environmental (10.2%) information processing. Pathways relating to cellular processes (7.2%) and organismal systems (2.7%) had some representation (Figure 3.8).

Importantly, a proportion of the ASVs submitted to Piphillin inferred functional pathways in human disease (6.3%). The pathways were related to several current concerns in the fresh produce food safety space, including antimicrobial drug resistance, and bacterial, parasitic, and viral infectious diseases (Figure 3.8). KEGG pathways implicated in Tuberculosis, Legionellosis and Pertussis were predicted from the samples, as well as infection with *Helicobacter pylori* and *Salmonella*. Pathways implicated in Shigellosis and infection with *Staphylococcus aureus* and *Vibrio cholerae* were also identified but were comparatively insubstantial (Figure 3.9).

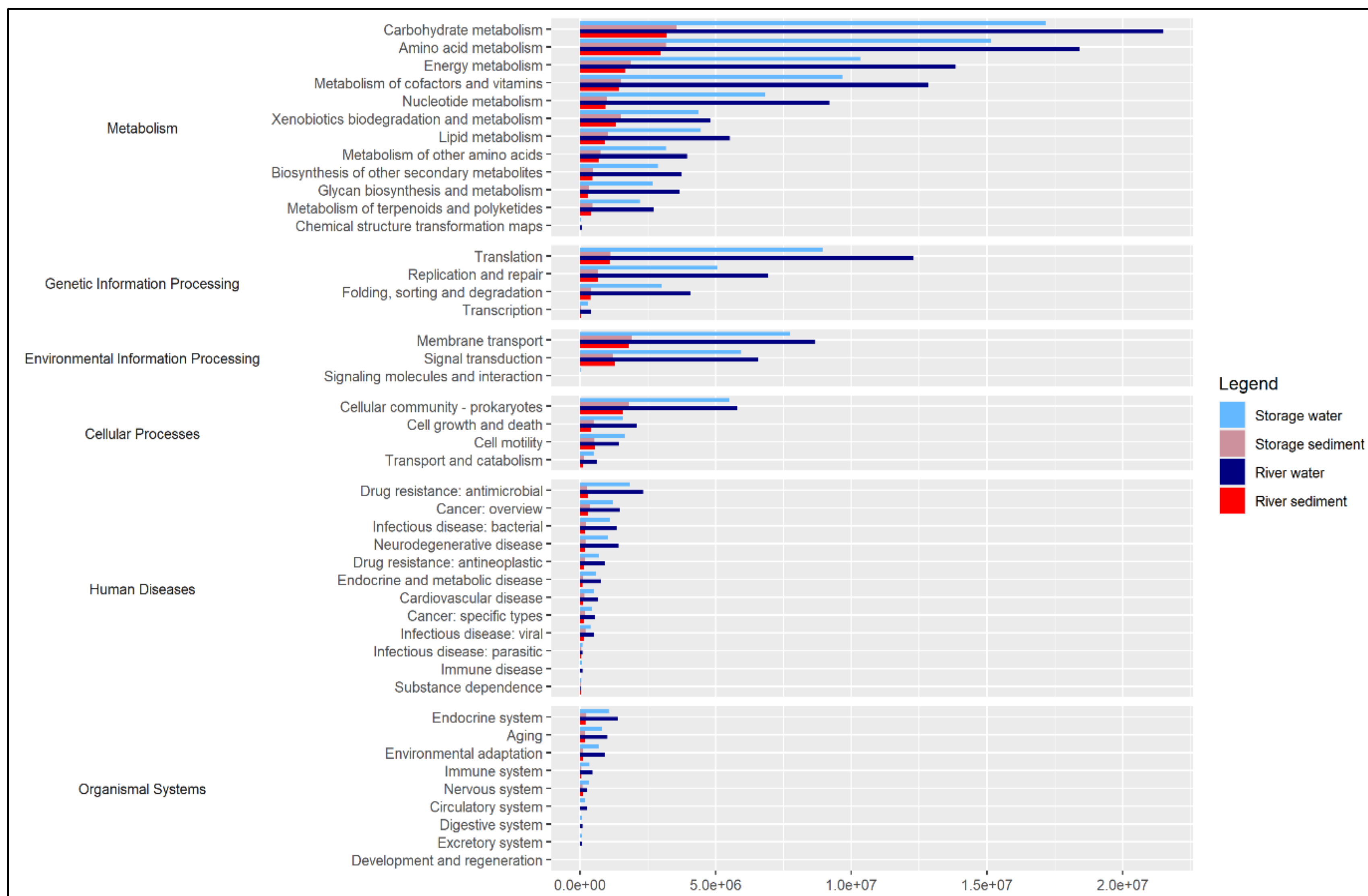


Figure 3.8: KEGG pathways predicted by Piphillin, separated by sample source and matrix.

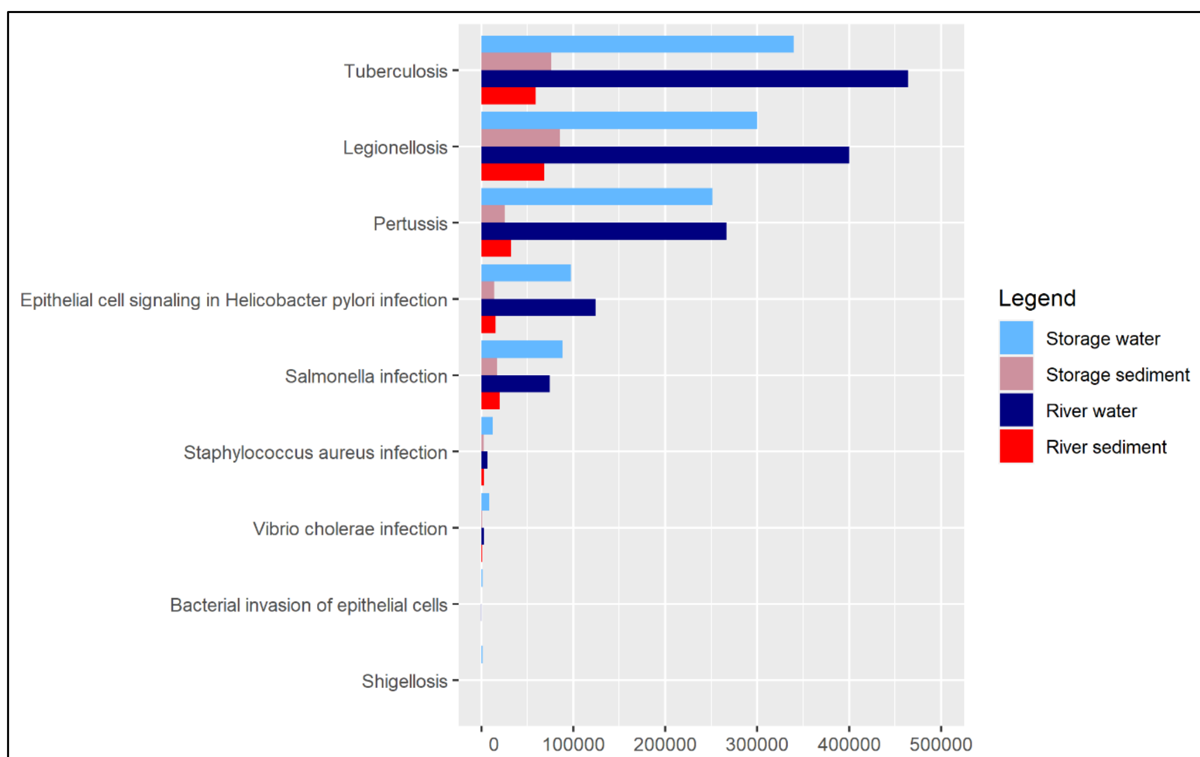


Figure 3.9: KEGG pathways predicted by Piphillin in the 'Infectious disease: bacterial' tier, separated by sample source and matrix.

3.4 Discussion

This chapter describes the bacterial community compositions and functional classes in water and sediments from the upper Hawkesbury River area. Across 190 samples, this study was able to generate 10,146,372 16S rRNA sequence reads, represented by 67,390 unique ASVs, a majority of which (66.1%) could be annotated to genus level. The results reveal that there is substantial diversity in the bacterial populations in these sources collectively, with the most considerable number of unique taxa in the sediment (52,321 ASVs) and within the Proteobacteria phylum (21,559 ASVs). The inferred functional profiles of the bacterial communities revealed associations with human diseases of relevance to the fresh produce industry, such as infection with *Legionella*, *Salmonella*, *Staphylococcus*, *Vibrio*, and *Shigella*. Using the relative abundance of selected genera as an indicator of food safety concerns, the results show that sediment, on-farm water storages, and warmer sampling seasons all lead to greater potential risk for foodborne pathogenic contamination.

3.4.1 Diversity

Comparing the average number of ASVs and hits per sample as an approximation of bacterial diversity, the results suggest that diversity was highest in summer and lowest in spring, while autumn and winter were comparable. Many published studies have found seasonal differences in bacterial diversity in environmental waters (Chen et al., 2018;

Roberto et al., 2018; Shang et al., 2022; Shilei et al., 2020), with some misalignment on whether diversity is higher in summer (Shilei et al., 2020) or winter (Roberto et al., 2018; Shang et al., 2022). It is unusual that the samples in the current study do not reflect the traditional summer-winter dichotomy, but only a single iteration of each season was sampled and therefore it may be that the sampling period did not reflect typical seasonal variation for the region. Though there were more ASVs and hits from the river sites than the on-farm storage sites, they are comparable in their diversity per sample, suggesting that the habitats are equally conducive to harboring complex bacterial communities, despite differences in physicochemical water quality (Table 2.2). In contrast, the sediment was revealed to have substantially more diversity than the water column, consistent with other freshwater studies (Abia et al., 2018; Staley and Sadowsky, 2016).

3.4.2 Taxonomic profiles

The irrigation water sources contained complex bacterial compositions, which were explored with human health implications in mind. At the phylum level of taxonomic rank, it can be concluded that irrigation water sources have a consistent community structure. Across all four seasons, within river and on-farm sites, and in both sediment and water, the key groupings were similar. The dominance of Proteobacteria and Bacteroidota in the irrigation water sources point to a possibility of enteric illness due to the pathogens those phyla contain, yet this is not an uncommon finding in surface water metagenomic studies (Filippini et al., 2019; Ghaju Shrestha et al., 2017; Nho et al., 2018; Staley and Sadowsky, 2016).

The Proteobacteria were examined more closely to evaluate the relative abundance of specific human health-related taxa. Proteobacteria, specifically Gammaproteobacteria, are important because they comprise several known human pathogens (Rizzatti et al., 2017), including the genera *Escherichia*, *Shigella*, *Salmonella*, and *Yersinia*. The most abundant Proteobacteria orders were not relevant to enteric infectious disease, but the less abundant Enterobacterales order provided crucial information on the detection of *Escherichia/Shigella*. Despite *E. coli* detections in every water column sample using a traditional, culture-based method (Table 2.2), the *Escherichia/Shigella* genus was almost exclusively detected in the sediment of on-farm storages, and primarily in winter, despite all samples undergoing the same metagenomic workflow. One simple explanation for this result could be that *Escherichia/Shigella* DNA from dead cells is persisting in these samples more so than in the other samples. This result may also be influenced by aspects of the metagenomic workflow, such as the strict quality filtering of the sequence reads, and the conversion of sequence reads to relative abundance overwhelming the detection. Otherwise, this discrepancy in results could be suggesting that on-farm storage sediments are ideal habitats for coliform

bacteria in a viable but nonculturable state (Servais et al., 2009), however, it is imprudent to make this conclusion using just the data generated in this chapter and without further study.

3.4.3 Irrigation pathogen potential index (IPPI)

The relative abundances of the genera identified as being irrigation water concerns, summarised as the IPPI, revealed that the risk of potential pathogens was usually higher in the sampled sediments than in the water sources. As highlighted in 2.4.1 above, sediment communities contribute significantly to bacterial communities in the water column (Benjamin et al., 2013; Bradshaw et al., 2016; Droppo et al., 2009; Haller et al., 2009; Staley et al., 2016), and therefore to pathogen loading. The potential pathogens present in sediment pose a risk when the water sources are disturbed and sediment-bound bacteria could be resuspended into the waterway (Sassi et al., 2020).

The IPPI was generally a small relative abundance, with the mean value being less than 1% of the overall community. However, this is not necessarily a negligible amount, as it has been found that some pathogens can initiate an infection with only a few cells (Schmid-Hempel and Frank, 2007). Even with a very limited number of ASVs identified to species level (0.4%), there were confirmed sequences belonging to *Staphylococcus aureus, Bacillus cereus, Campylobacter coli, Enterococcus faecium, Escherichia/Shigella coli,* and *Listeria monocytogenes* that were incorporated into the IPPI. Due to restraints in the depth of sequencing currently available for 16S rRNA, it was not possible to determine the strain of these potential pathogens, not the presence of virulence genes. However, if used as an indicator of microbial risk, the IPPI appears to be a useful tool.

The sensitivity and value of the IPPI could be increased if the index discriminated between *Pseudomonas* species or subgroups. The *Pseudomonas* portion of the IPPI was dominant and therefore the most influential; *Pseudomonas* was the most abundant genus within the index regardless of sample source and matrix (Table 3.4) or season (Table 3.5). This may be problematic because the IPPI likely overestimates the abundance of harmful or pathogenic *Pseudomonas* species or subgroups. A potential solution to this issue and an improvement in the utility of the IPPI could be to limit the inclusion of *Pseudomonas* in the index to those species known to be pathogenic, such as *P. aeruginosa* (Streeter and Katouli, 2016). This will be possible with improved species-level metagenomic identification.

3.4.4 Functional profiles

As the 16S rRNA gene cannot reveal a microbe's functional capacity directly, the functions of the bacterial communities were inferred using Piphillin, revealing that the sampled bacterial communities are involved in diverse functional pathways. It is important to keep in

mind when interpreting the functional profiles that only 481 of the 10,000 ASVs (<5%) submitted to Piphillin had greater than 99%-identity to sequences in the KEGG database and were included in the analysis. From this small subset of the bacterial community, a few generalisations can be made.

Firstly, most of the KEGG pathways identified were related to metabolism, which is a common finding in similar studies (Abia et al., 2018; Staley et al., 2016). Secondly, the communities had similar allocations of functions regardless of whether they were sampled from the rivers or on-farm. This suggests that a community's function is primarily dictated by its habitat, i.e., water or sediment, which is mirrored in the taxonomic profiles and supported by existing literature (Staley et al., 2016).

Lastly, functional pathways involved in antimicrobial drug resistance and human infectious diseases were inferred in the samples, including infection with the enteric pathogens *Salmonella*, *Vibrio*, and *Shigella*. This result appears to support the conclusion that the irrigation water sources are capable of harboring potential human health risks. However, the functional profiles should be interpreted with caution because there are several assumptions that underpin the inferences by Piphillin. Piphillin pairs sampled 16S rRNA sequences directly with sequences from whole genomes in the KEGG database, using nearest-neighbour matching to infer the closest genome (Iwai et al., 2016). The functions of the genomes as annotated in the KEGG database are then used to extrapolate the functions of the samples, after adjusting for copy number. The constraint with this approach is that 16S rRNA sequences are usually only accurate to the genus level, and microbial functions are determined at the strain level (Garner et al., 2021). Further, it has been suggested that only deep sequencing can achieve accurate function profiling (Knight et al., 2012). This means that while there was genetic material associated with enteric pathogens and antimicrobial resistance highlighted by Piphillin, it should not be considered a reliable detection.

3.4.5 Limitations

Several aspects of the chosen molecular analysis methodology will have impacted the results of this chapter and will consequently also influence Chapter 4 below. The method of DNA concentration and extraction used for this study was a customised filtration process and an industry-standard extraction kit. There is not a consistent approach among metagenomic studies as to which method of DNA extraction is used, even when comparing only 16S studies of water (Lear et al., 2018), however, the PowerSoil Pro kits have been shown to retain more microbial diversity through metagenomic workflows than in-house methods (Mateus-Barros et al., 2019) and were considered more accessible for use in this study than methods that might have required validation or specialist knowledge. The V1-V3

hypervariable region of the 16S gene was chosen as the target for this study as it was in keeping with the Australian Microbiome Initiative. The 16S gene is over 1550 base pairs long (Clarridge, 2004), which exceeds the current Illumina MiSeq platform maximum of 300 base pairs, meaning whole genomes are not sequenced using this technology. The choice of which of the nine hypervariable regions to sequence could have affected the precision or diversity when the taxonomies were aligned (Guo et al., 2013). However, the DADA2 pipeline is robust, and the quality filtering parameters were set to balance the requirement for high-quality ASVs with the need for ample representation of the diversity of taxa. Rarefaction curves (Figure 7.3) showed that the filtering parameters were sufficient; further reads would not likely have resulted in the generation of more ASVs.

The taxonomic database was also chosen after considering the potential impacts. The SILVA database was preferred because it worked well with the DADA2 pipeline and the Piphilin server. However, it is important to note that the most common taxonomic databases (i.e., SILVA, Greengenes, RDP, NCBI, OTT) vary substantially in the taxa they contain (Balvočiute and Huson, 2017), which could lead to incomplete identification of the bacterial profiles in this study. It would be possible in the future to attempt to construct a food safety-specific taxonomic database to achieve better species-level assignment (F Escapa et al., 2020). There are databases targeting specific genera such as Cyanobacteria (CyanoDB, Hauer and Komárek, 2022) and targeting specific habitats such as freshwater lakes (TaxAss, Rohwer et al., 2018).

Piphilin was selected as the functional prediction tool because it was shown to perform better than alternative tools PICRUSt (Langille et al., 2013) and Tax4Fun (Aßhauer et al., 2015) when used on human clinical samples (Iwai et al., 2016). The Piphilin server by default uses the KEGG genomic database but the current alternative is BioCyc (Karp et al., 2019). A study comparing both found that KEGG pathways are larger than BioCyc pathways and are therefore less likely to be conserved biological processes (Green and Karp, 2006). Using high-quality ASVs over OTUs and specifying a 99%-identity threshold in the functional inference has produced results that are likely reliable but insufficient for a more thorough analysis. However, this stringency does not overcome the key limitation of functional inference, which is that the functionality is predicted based on taxonomy rather than directly examining functional genes (Garner et al., 2021). The functional profiles presented in this chapter should point to useful areas of follow-up studies rather than being considered as accurate.

However, the ultimate limitation of using 16S rRNA to generate taxonomic and functional profiles as it relates to irrigation water sources is that it is extremely difficult to convert the

results to a measure of actual human-health risk. Most bacteria have multiple copies of the 16S rRNA gene (Větrovský and Baldrian, 2013), which means that knowing the correct copy number in each identified species is important. This was not addressed in this thesis. Further, a vast majority of the ASVs in this study could only be identified to the family or genus level; while this remains an impressive feat for bacterial communities that are largely unculturable, it cannot be confirmed that viable, pathogenic strains were present at the time of sampling, meaning the actual human health risk is unquantified. However, this limitation also applies to the current indicator of water quality risk, as generic *E. coli* is not sub-typed as a matter of standard monitoring procedures. There is a possibility of using NGS techniques for the absolute quantification of pathogenic species in the future as the technology advances, and these would need stringent internal controls like in qPCR.

3.5 Conclusion

The NGS analysis of the collected water and sediment samples successfully identified fresh produce food safety risks in the irrigation water sources. Specifically, *Bacillus cereus*, *Campylobacter coli*, *Listeria monocytogenes*, *Staphylococcus aureus*, *Enterococcus faecium,* and *Escherichia/Shigella coli* were detected. The irrigation pathogen potential index was developed and used to demonstrate that the risk of pathogens was higher in summer and was associated with on-farm storage sediments. Overall, this chapter enhances the previous chapter's conclusion that human health risks are persistently present in irrigation water sources.

4 Dynamics of bacterial populations and human pathogens in irrigation water sources

4.1 Introduction

Next generation sequencing (NGS) technologies can reveal a substantially larger proportion of microbial communities than traditional, culture-dependent techniques. The benefit of this is that investigations into the microbial diversity of natural environments can explore in detail not just the identities and abundances of the bacteria present, but their roles in sustaining community stability, diversity, and functioning. Through such studies, we can determine in extrinsic and intrinsic drivers of bacterial community composition (Ibekwe et al., 2016; Nevers et al., 2020; Wang et al., 2018).

A common analysis of microbial communities is to quantify the alpha (in-sample) and beta (between sample) diversity (Chen et al., 2018; Roberto et al., 2018; Shang et al., 2022; Shilei et al., 2020). Zhang et al. (2021) found that sediment bacterial communities are more diverse than water communities, and that they are driven by contrasting environmental factors, which has also been reported in other studies (Sun et al., 2019). The authors concluded that water bacterial diversity was more susceptible to climate effects, while sediment bacterial diversity was driven primarily by physicochemical parameters (Zhang et al., 2021).

The construction of microbial networks can be used for a variety of purposes, including predicting hub species and species interactions (Röttjers and Faust, 2018), identifying candidate microbes for disease management (Poudel et al., 2016), and monitoring ecosystem change (Barroso-Bergadà et al., 2021). These networks can be quantified by calculating such measures as community connectedness and cohesion (Herren and McMahon, 2017). Some studies have identified specific keystone taxa, which are taxa with a disproportionately substantial role in maintaining community stability within these networks (Dai et al., 2019; Herren and McMahon, 2018). For example, Graham et al. (2017) identified two keystone families in nearshore groundwater samples (one family of unassigned organisms and one belonging to the candidate phylum *OP3*) and two keystone families in inland groundwater samples (*Chloracidobacteria* and *Chloroflexi*).

This chapter aims to explore how bacterial dynamics relate to water quality and safety through three objectives: (1) explore the alpha and beta diversity of the collected water and sediment samples, (2) calculate the connectedness and cohesion measures from the defined bacterial communities in these samples, and (3) evaluate the statistical correlation of

the concentration of *E. coli* to the developed irrigation pathogen potential index and other bacterial community characteristics of the sampled irrigation water sources.

4.2 Methods

4.2.1 Sampling and previous analyses

The samples and accompanying data for this chapter were as described previously (see 2.2 and 3.2 above). Water and sediment samples were collected monthly from the upper Hawkesbury River area in Greater Sydney, New South Wales, from September 2018 to August 2019. In total, 190 samples for analysis (96 water, 94 sediment) were collected from 8 sites (5 river, 3 on-farm storage) over a 12-month period. The accompanying data included *Escherichia coli* (*E. coli*) concentrations for the water and sediment, physicochemical water quality, meteorological conditions, and sediment particle size determinations. Metagenomic analyses of the samples using the DADA2 pipeline (Callahan et al., 2016a, 2016b) resulted in 67,390 amplicon sequence variants (ASVs) for the 16S rRNA V1-V3 hypervariable region. The SILVA database (McLaren and Callahan, 2021b; Quast et al., 2013) and Piphillin server (Iwai et al., 2016; Narayan et al., 2020) were used to construct the taxonomic and functional profiles of the bacterial communities, respectively.

All statistical analyses were performed in R (R Core Team, 2021).

4.2.2 Alpha diversity

The Shannon diversity index was used to quantify alpha diversity (Shannon, 1948). All 67,390 ASVs were included in the calculation. Diversity was calculated for each sample using the ‘phyloseq’ package (McMurdie and Holmes, 2013). Significant ($P < 0.05$) pairwise comparisons between the samples were determined using Wilcoxon rank sum tests (Wilcoxon, 1945).

4.2.3 Beta diversity and PERMANOVA

Ordination was used to examine the similarities and differences of the sampled bacterial communities beyond the direct comparisons of the taxonomic and functional profiles as previously described (see section 3.3). The rank-based approach of nonmetric multidimensional scaling (NMDS) was chosen as the ordination method due to its suitability for beta diversity or community distance matrices, such as Bray-Curtis (Kruskal and Wish, 1978).

Due to the very large number of ASVs generated in this study, it was necessary to include a subset to perform and visualise the ordinations in a meaningful way. ASVs were filtered to

only those that appeared a minimum of ten times in a minimum of 10% of the samples; those that were identified to at least the genus level of taxonomic rank; and those that belonged to the five most abundant phyla.

The 'vegan' package (Oksanen et al., 2022) was used to calculate the Bray-Curtis distances and perform the NMDS ordinations. Ordination was performed for all the samples (n = 190), then the water (n = 96) and sediment (n = 94) samples separately. Each time, the ordination was set to a maximum of 1,000 random starts to find a stable solution. Stress plots were used to select the number of dimensions for each NMDS, with stress values less than 0.05 considered optimal. The 'envfit' function was used to project environmental and water quality parameters onto the water and sediment NMDS plots as extrinsic variables. The parameters were: antecedent rainfall (mm, discount factor = 0.7) and temperature (°C, discount factor = 0.7), pH, turbidity (NTU), conductivity ($\mu S\ cm^{-1}$), dissolved oxygen (DO; % saturated), chlorophyll *a* ($\mu g\ L^{-1}$), oxidative-reductive potential (ORP) (mV), phycocyanin ($\mu g\ L^{-1}$), fluorescent dissolved organic matter (fDOM) (RFU), and alkalinity (as $CaCO_3$ mg L^{-1}).

Permutational multivariate analysis of variance (PERMANOVA) was used to statistically examine the variation in the bacterial communities in response to the environmental and water quality parameters. The 'adonis2' function was used to perform the PERMANOVAs on the Bray-Curtis distances for the water and sediment samples separately. The function was set to 1,000 permutations and the significance was assessed for the marginal effects.

4.2.4 Connectedness and cohesion

Connectedness for each ASV, and positive and negative cohesion for each sample were calculated according to the methods outlined in Herren and McMahon (2017). A minimum ASV persistence of 5% was first used to subset the ASVs to a number that was computationally achievable and to reduce the chance of spurious correlations for rare taxa (Herren and McMahon, 2017). Positive and negative connectedness were calculated as ASV-to-ASV correlations with an in-built null model correction (Herren and McMahon, 2017). Positive and negative cohesion were calculated by multiplying the relative abundance of the ASVs within each sample by their associated positive and negative connectedness. The cohesion metric is therefore bounded by -1 to 0 for negative cohesion and 0 to 1 for positive cohesion. The 'taxon shuffle' algorithm for the null model as defined by the authors (Herren and McMahon, 2017) was used for this analysis, with 200 iterations. Significant ($P < 0.05$) pairwise comparisons between the samples were determined using Wilcoxon rank sum tests (Wilcoxon, 1945).

4.2.5 *E. coli* and community characteristics

Statistical comparisons across the water and sediment samples were used to evaluate whether generic *E. coli* was strongly correlated to the bacterial community characteristics. Pairwise Spearman correlations were calculated using the 'stats' package in R (R Core Team, 2021) for *E. coli* in the water and sediment as estimated by IDEXX Colilert, the relative abundance of the order Enterobacterales, the irrigation pathogen potential index, Shannon (alpha) diversity, and the community cohesion measures.

4.3 Results

4.3.1 Alpha diversity

The Shannon (alpha) diversity of the samples ranged from 2.69 to 7.42, with a mean value of 5.58. Overall, the sediment samples had significantly higher Shannon diversity than the water samples (P < 0.0001) (Figure 4.1). The median Shannon diversities were highest for the sediment samples (storage sediment 6.70, and river sediment 6.57), which were not significantly different between source types (P > 0.05). The median Shannon diversities were lowest for the water samples (storage water 4.65, and river water 4.83), which were also not significantly different between source types (P > 0.05).

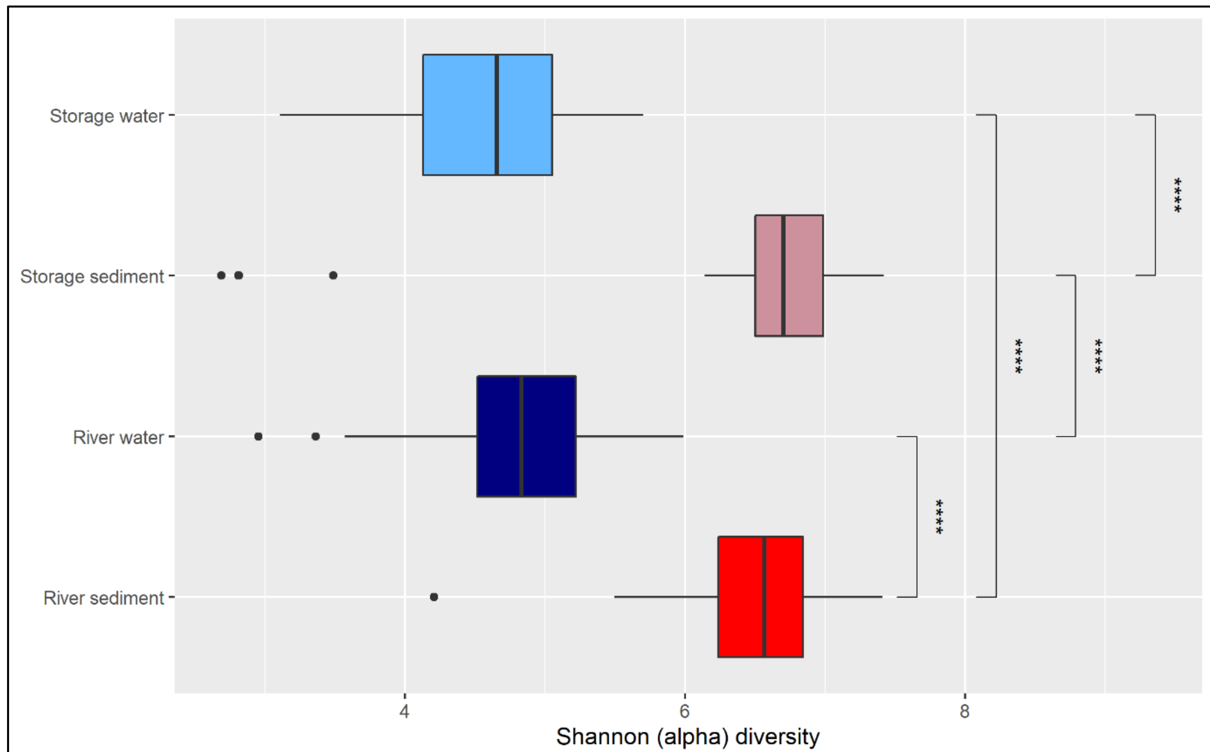


Figure 4.1: Shannon (alpha) diversity by sample matrix and source type. Asterisks are indicative of pairwise significance (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown).

Shannon diversity (Figure 4.2) was highest in Autumn (March to May) with a median value of 5.85, but it was not significantly higher than in Summer (December to February) which had a median value of 5.62. Shannon diversity was lower in Winter (June to August), and significantly lower ($P < 0.05$) in Spring (September to November) and with median values of 5.09 and 5.34, respectively.

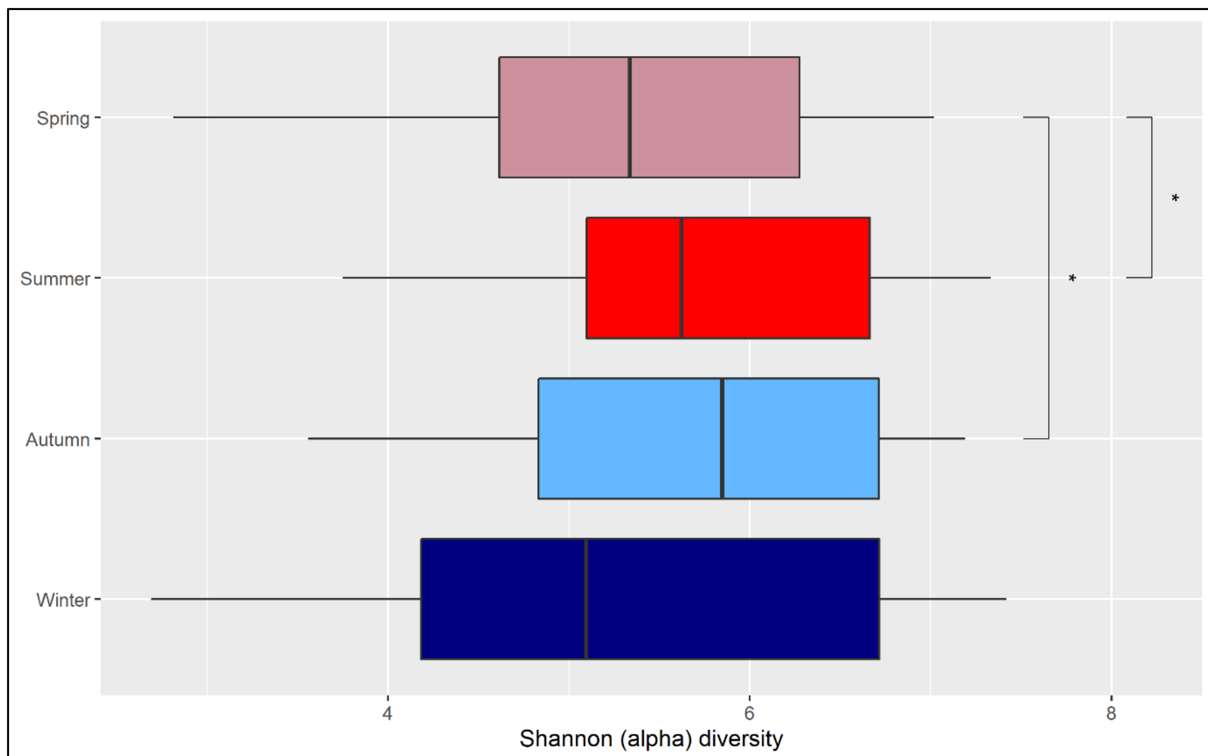


Figure 4.2: Shannon (alpha) diversity by season (Spring: September-November; Summer: December-February; Autumn: March-May; Winter: June-August). Asterisks are indicative of pairwise significance (**** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, $P > 0.05$ not shown).

4.3.2 Beta diversity and PERMANOVA

A subset of 418 ASVs were selected for ordination and visualisation after meeting the criteria described: appeared a minimum of ten times in a minimum of 10% of the samples, was identified to at least the genus level of taxonomic rank and belonged to the five most abundant phyla (i.e., Proteobacteria, Actinobacteria, Bacteroidota, Cyanobacteria, and Chloroflexi). This subset represented less than 1% of the total number of ASVs but represented a mean relative abundance of 31.4% per sample. The NMDS had a satisfactory stress value of 0.16 when fit with 10 dimensions. Figure 4.3 shows the distribution of the samples across the most informative two axes. The samples are coloured by the water source type (i.e., river and on-farm) and the sample matrix (i.e., water and sediment), and the ellipses are indicative of that grouping. The most notable result is the demarcation between the water communities and the sediment communities along the first axis; excluding a small number of outliers, the water and sediment samples are well-segregated. This was

confirmed with a PERMANOVA on the Bray-Curtis distances for the sample types, showing that the locations of the centroids are significantly different (F = 1.72; P < 0.01). Also suggested by the ellipses in the NMDS plot were the closer groupings of the river samples compared to the on-farm storage samples along the second axis. However, this was rejected by an analysis of the homogeneity of the dispersion, showing that the dispersions were not significantly different (F = 1.22; P = 0.311).

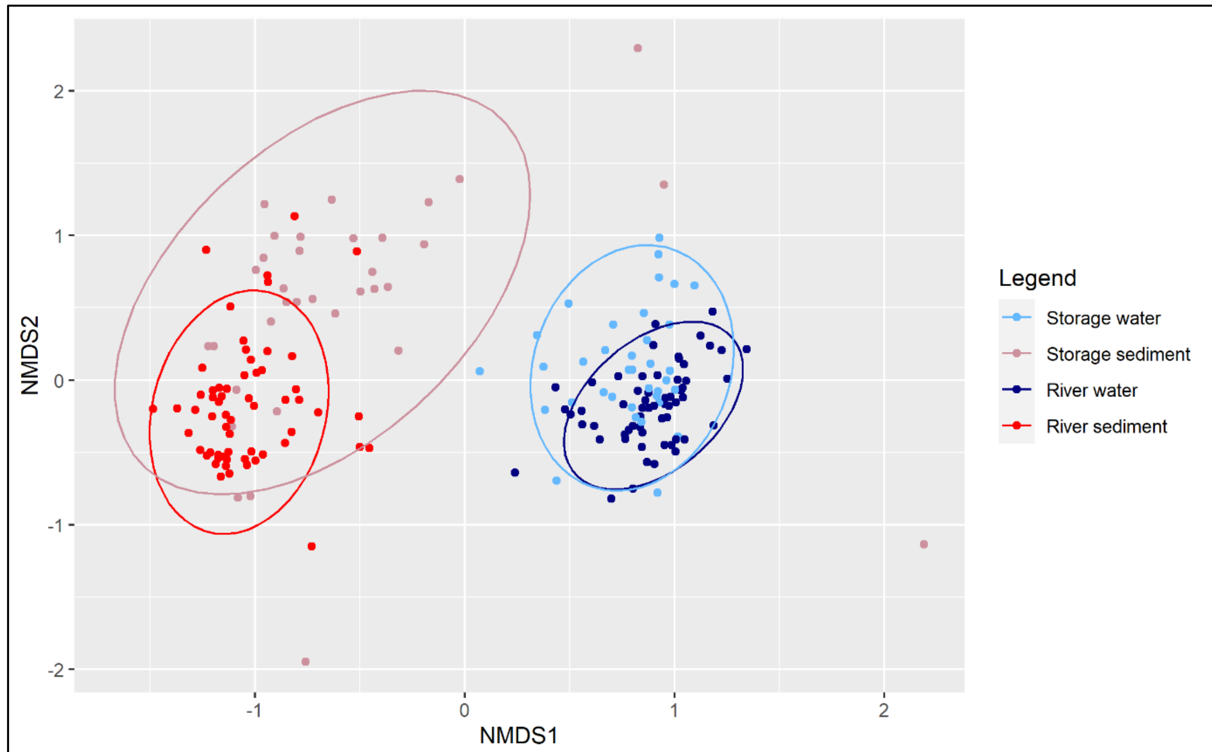


Figure 4.3: Nonmetric multidimensional scaling (stress = 0.16) of the water and sediment samples (n = 190). The ellipses are indicative of group distributions only. PERMANOVA on the Bray-Curtis distance for the sample types was significant (P < 0.01). Analysis of the homogeneity of the dispersion for the sample types was not significant (F = 1.22; P = 0.311).

The NMDS of the water samples (n = 96) had an excellent stress value of 0.048 when fit to ten dimensions. The NMDS plot revealed that community variation within the water samples was correlated with the meteorological and physicochemical parameters, as shown by the relative length of the arrows projected onto the ordination (Figure 4.4). The PERMANOVA confirmed that positive community cohesion, IPPI, Shannon diversity, sediment clay content, rainfall, temperature, solar exposure, pH, turbidity, conductivity, DO, ORP, fDOM, alkalinity, chlorophyll *a*, phycocyanin, and relative abundance of Enterobacterales were significant explanatory variables for the community assemblages (P < 0.05) (Table 4.1).

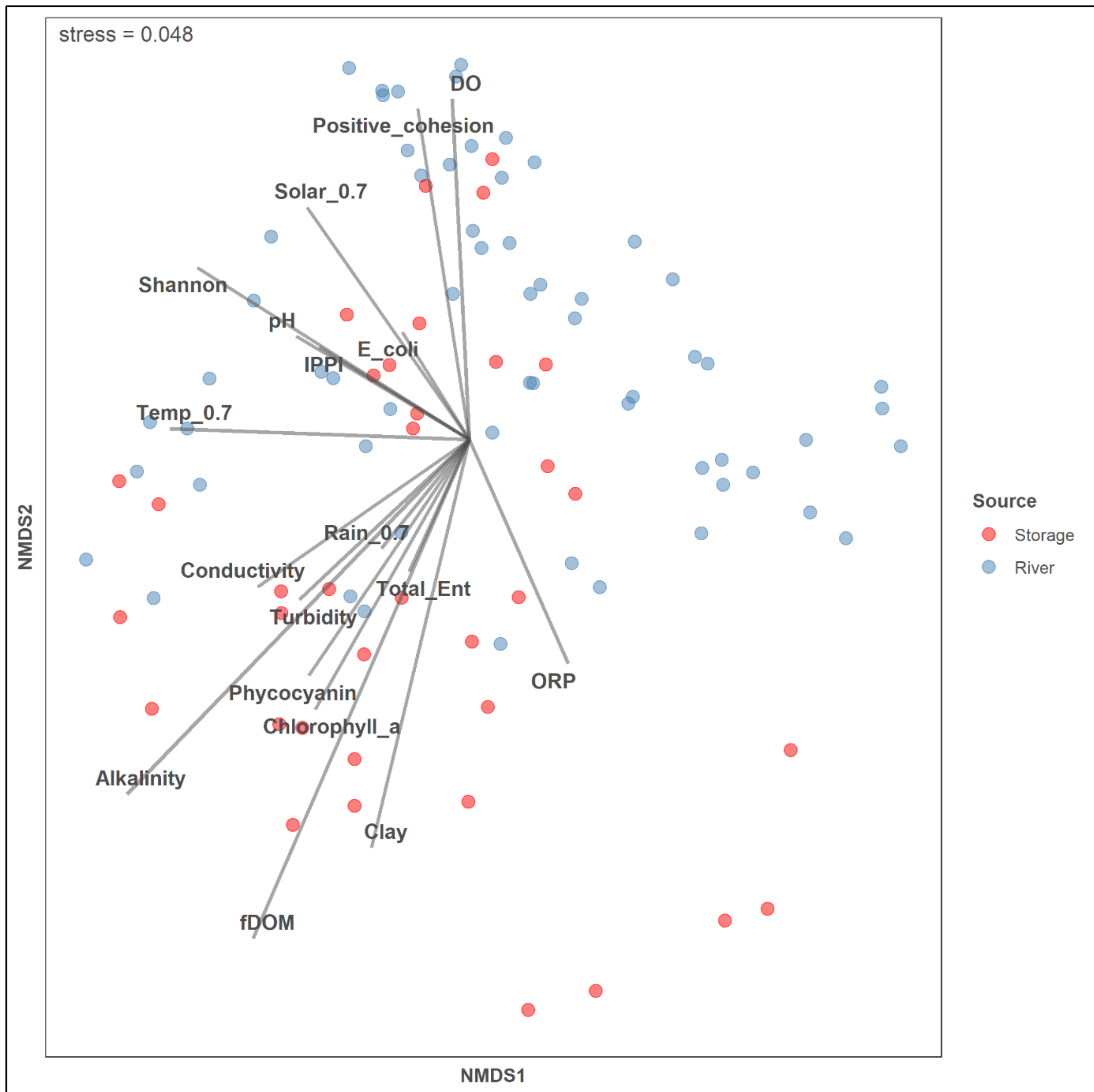


Figure 4.4: Nonmetric multidimensional scaling (stress = 0.048) of the water samples (n = 96). The arrows are indicative of a linear trend with the ordination scores only.

Table 4.1: PERMANOVA output for the Bray-Curtis distances of the water samples (n = 96) showing the marginal effect of terms with NA values excluded. Asterisks are indicative of significance (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown).

	Df	SumOfSqs	R2	F	P	Significance
Positive cohesion	1	1.2254	0.04675	6.3459	0.001	***
IPPI	1	0.6634	0.02531	3.4357	0.001	***
Shannon	1	0.6172	0.02355	3.1962	0.002	**
Clay	1	0.4198	0.01601	2.1739	0.008	**
Rain_0.7	1	0.7237	0.02761	3.7475	0.001	***
Temp_0.7	1	0.6249	0.02384	3.2359	0.001	***
Solar_0.7	1	0.5355	0.02043	2.7728	0.002	**
pH	1	0.3793	0.01447	1.9643	0.011	*
Turbidity	1	0.3776	0.01441	1.9556	0.018	*
Conductivity	1	0.5481	0.02091	2.8385	0.001	***
ORP	1	0.3139	0.01197	1.6254	0.05	*
fDOM	1	0.5204	0.01985	2.695	0.003	**
Alkalinity	1	0.507	0.01934	2.6257	0.003	**
Phycocyanin	1	0.332	0.01267	1.7194	0.023	*
Chlorophyll *a*	1	0.3496	0.01334	1.8106	0.018	*
Enterobacterales	1	0.4209	0.01606	2.1795	0.002	**
Residual	71	13.7106	0.52302			
Total	87	26.2141	1			

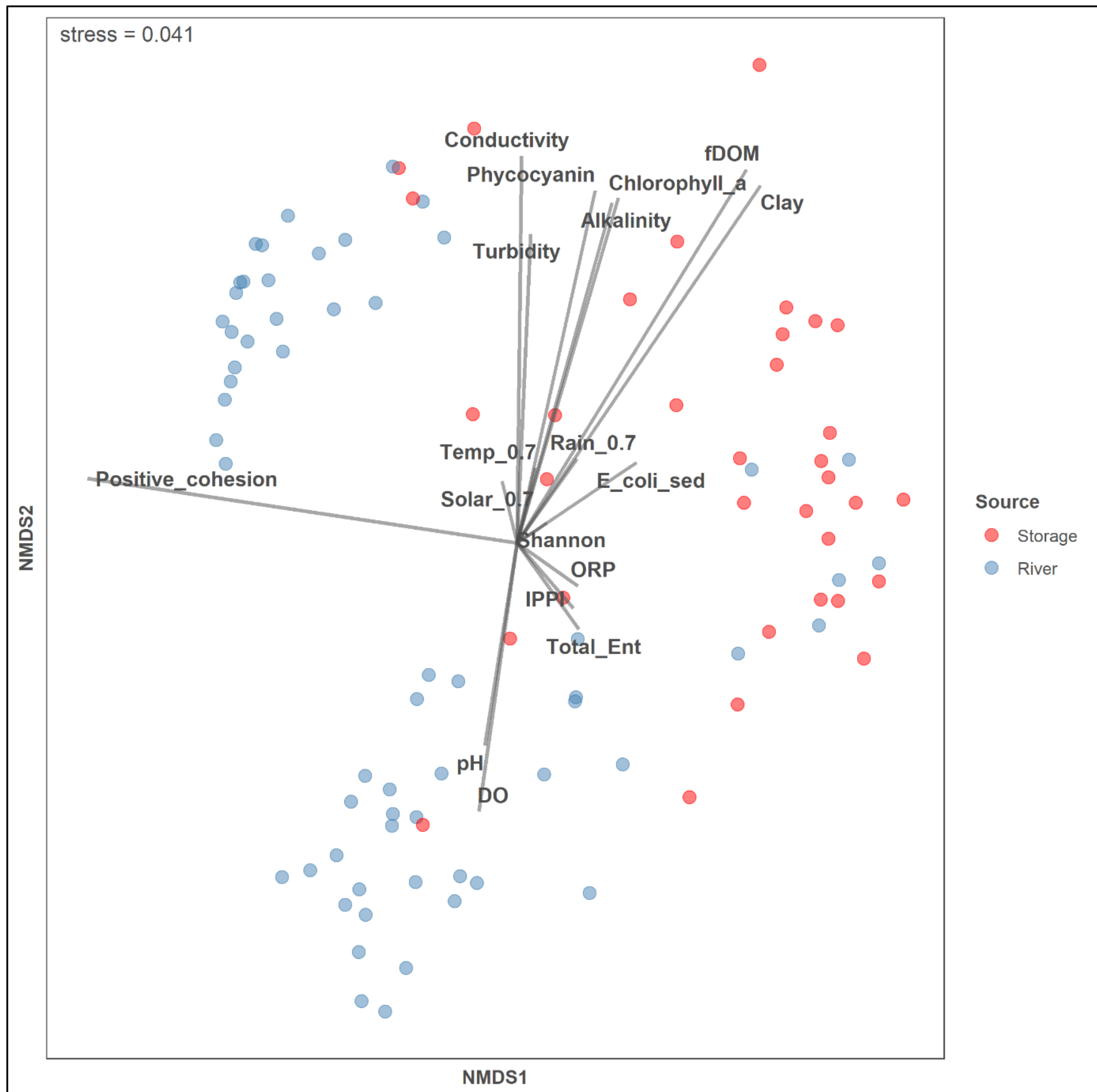


Figure 4.5: Nonmetric multidimensional scaling (stress = 0.041) of the sediment samples (n = 94). The arrows are indicative of a linear trend with the ordination scores only.

The NMDS of the sediment samples (n = 94) also had an excellent stress value of 0.041 when fit to ten dimensions. The NMDS plot suggests that community variation within the sediment samples was correlated with physicochemical parameters and somewhat with the meteorological parameters (Figure 4.5), as shown by the relative length of the arrows projected onto the ordination. The NMDS plot shows some segregation of the storage and river sediment samples. The plot also appears to cluster the river samples into two distinct groups, with increasing pH and DO both associated with the group on the bottom of the plot, and positive community cohesion associated with the group on the top left. The PERMANOVA confirmed that positive community cohesion, IPPI, Shannon diversity, sediment clay content, pH, ORP, fDOM, and relative abundance of Enterobacterales were

significant explanatory variables for the variation in sediment community assemblages (P < 0.05) (Table 4.2).

Table 4.2: PERMANOVA output for the Bray-Curtis distances of the sediment samples (n = 94) showing the marginal effects with NA values excluded. Asterisks are indicative of significance (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown).

	Df	SumOfSqs	R2	F	P	Significance
Positive cohesion	1	3.5478	0.11357	13.3733	0.001	***
IPPI	1	0.5374	0.0172	2.0258	0.004	**
Shannon	1	0.7003	0.02242	2.6397	0.002	**
Clay	1	0.8458	0.02707	3.1881	0.001	***
pH	1	0.6615	0.02118	2.4935	0.003	**
ORP	1	0.9872	0.0316	3.7212	0.001	***
fDOM	1	0.5476	0.01753	2.0643	0.007	**
Enterobacterales	1	0.4671	0.01495	1.7608	0.008	**
Residual	77	20.4275	0.65391			
Total	85	31.2389	1			

4.3.3 Connectedness and cohesion

A subset of 2,218 ASVs were selected for the connectedness and cohesion calculations after meeting the criterion of being prevalent in at least 5% of the samples. This subset represented 3.3% of the total number of ASVs but represented a mean relative abundance of 55.8% per sample. The positive connectedness values ranged from 0.015 to 0.366 with a mean value of 0.178. The taxonomies of the ten ASVs with the highest positive connectedness are shown in Table 4.3. Eight of these top ten taxa were Proteobacteria, one was a Cyanobacteria, and one was a Nitrospirota. The negative connectedness values reflected weaker correlations than the positive connectedness values, ranging from -0.003 to -0.155 with a mean value of -0.054. The taxonomies of the ten ASVs with the highest negative connectedness are shown in Table 4.4. Of these top ten taxa, four were Proteobacteria, four were Actinobacteriota, one was a Bacteroidota, and one was a Nitrospirota.

Table 4.3: The taxonomies of the ten ASVs with the highest positive connectedness

Phylum	Class	Order	Family	Genus	Positive connectedness
Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	0.366
Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Pedomicrobium	0.348
Proteobacteria	Alphaproteobacteria	Rhizobiales	A0839	NA	0.348
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Nitratireductor	0.345
Proteobacteria	Gammaproteobacteria	Cellvibrionales	Spongiibacteraceae	BD1-7 clade	0.341
Proteobacteria	Gammaproteobacteria	Burkholderiales	Comamonadaceae	Rubrivivax	0.341
Proteobacteria	Gammaproteobacteria	Burkholderiales	Nitrosomonadaceae	mle1-7	0.340
Cyanobacteria	Cyanobacteriia	Oxyphotobacteria Incertae Sedis	Unknown Family	Calothrix KVSF5	0.340
Nitrospirota	Nitrospiria	Nitrospirales	Nitrospiraceae	Nitrospira	0.338
Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	NA	0.334

Table 4.4: The taxonomies of the ten ASVs with the highest negative connectedness

Phylum	Class	Order	Family	Genus	Negative connectedness
Actinobacteriota	Actinobacteria	Frankiales	Sporichthyaceae	Candidatus Planktophila	-0.155
Actinobacteriota	Actinobacteria	Frankiales	Sporichthyaceae	hgcI clade	-0.138
Bacteroidota	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium	-0.135
Actinobacteriota	Actinobacteria	Frankiales	Sporichthyaceae	hgcI clade	-0.133
Proteobacteria	Gammaproteobacteria	Burkholderiales	Comamonadaceae	Limnohabitans	-0.125
Proteobacteria	Alphaproteobacteria	SAR11 clade	Clade III	NA	-0.125
Proteobacteria	Gammaproteobacteria	Burkholderiales	Methylophilaceae	Candidatus Methylopumilus	-0.122
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiales Incertae Sedis	NA	-0.117
Actinobacteriota	Actinobacteria	Frankiales	Sporichthyaceae	Candidatus Planktophila	-0.116
Nitrospirota	Nitrospiria	Nitrospirales	Nitrospiraceae	Nitrospira	-0.116

The positive cohesion values (i.e., positive ASV connectedness values multiplied by their relative abundance) were significantly different (P < 0.05) across all possible combinations of the source types and sample matrices (Figure 4.6). Median positive cohesion was highest in the river water (0.112), then the storage water (0.074), the river sediment (0.056), and lastly the storage sediment (0.009).

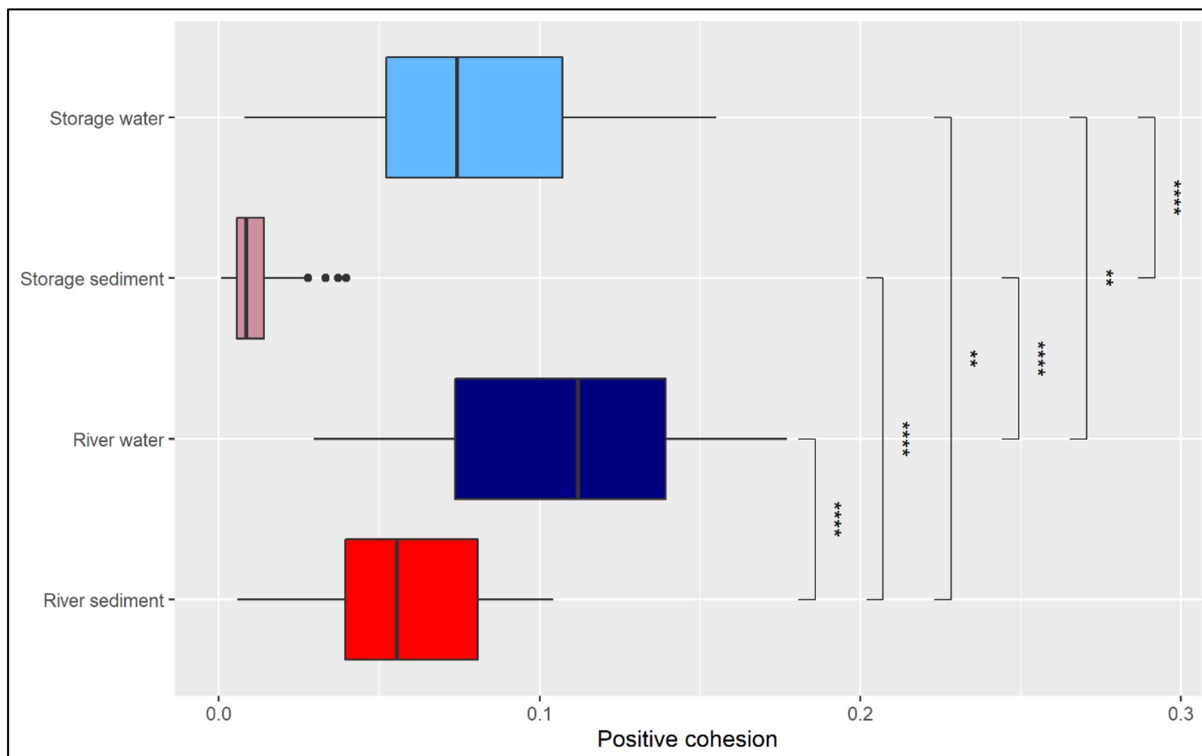


Figure 4.6: Positive cohesion values by sample matrix and source type. Asterisks are indicative of significant differences in the positive cohesion values as calculated by a Wilcoxon test (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown).

The negative cohesion values (i.e., negative ASV connectedness values multiplied by their relative abundance) were also significantly different (P < 0.05) across all possible combinations of the source types and sample matrices (Figure 4.7). Median negative cohesion was highest in the river water (-0.049), then the storage water (-0.031), the river sediment (-0.015), and lastly the storage sediment (-0.003).

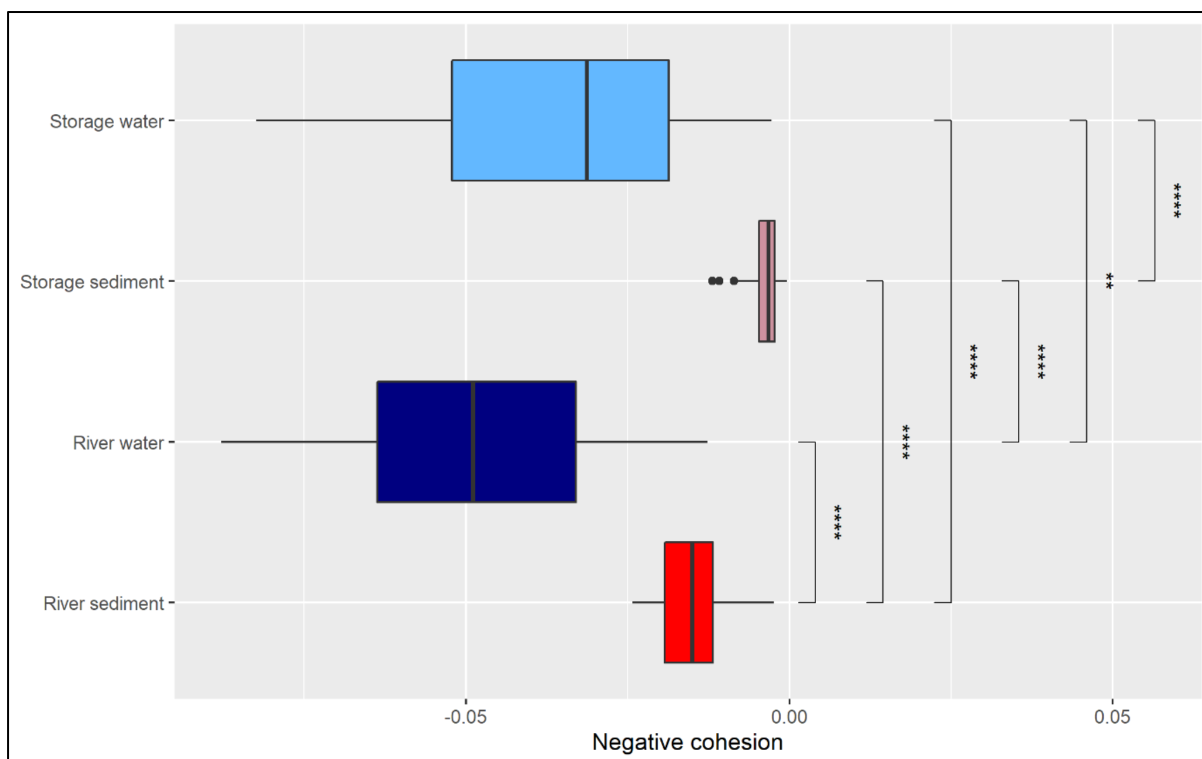


Figure 4.7: Negative cohesion values by sample matrix and source type. Asterisks are indicative of significant differences in the positive cohesion values as calculated by a Wilcoxon test (**** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05, P > 0.05 not shown).

A similar analysis was completed for the variation in positive and negative cohesion across the seasons, but no significant comparisons were discovered.

4.3.4 *E. coli* and community characteristics

The correlation analysis for the water samples (Table 4.5) showed that *E. coli* as estimated by IDEXX Colilert was not strongly correlated to the metagenomically-derived indices, except for a weak correlation (0.268) with Shannon diversity. Positive cohesion and negative cohesion were very strongly correlated to each other (-0.983), and substantially correlated to the IPPI (-0.533 and 0.526, respectively). There was also an expected, positive correlation between the IPPI and the relative abundance of the Enterobacterales order (0.326).

Table 4.5: Pairwise Spearman correlations for the water samples (n = 96).

	***E. coli* MPN (water)**	**Enterobacterales (% abund.)**	**IPPI**	**Shannon diversity**	**Positive cohesion**	**Negative cohesion**
***E. coli* MPN (water)**	1					
Enterobacterales (% abund.)	0.114	1				
IPPI	0.134	0.326	1			
Shannon diversity	0.268	0.112	0.262	1		
Positive cohesion	-0.093	-0.095	-0.533	-0.009	1	
Negative cohesion	0.063	0.105	0.526	0.041	-0.983	1

Table 4.6: Pairwise Spearman correlations for the sediment samples (n = 94).

	***E. coli* MPN (sediment)**	**Enterobacterales (% abund.)**	**IPPI**	**Shannon diversity**	**Positive cohesion**	**Negative cohesion**
***E. coli* MPN (sediment)**	1					
Enterobacterales (% abund.)	0.039	1				
IPPI	0.232	0.363	1			
Shannon diversity	0.229	0.073	-0.049	1		
Positive cohesion	-0.074	-0.388	-0.356	-0.290	1	
Negative cohesion	0.060	0.384	0.365	0.263	-0.991	1

The correlation analysis for the sediment samples (Table 4.6) also showed that *E. coli* in this study was not strongly correlated to the metagenomically-derived indices, except for weak correlation with Shannon diversity (0.229) and the irrigation pathogen potential index (0.232). However, amongst the indices, there was reasonable correlation (i.e., < -0.3, > 0.3) between the relative abundance of Enterobacterales, the IPPI, and positive and negative cohesion. Positive cohesion and negative cohesion were again very strongly correlated to each other (-0.991).

4.4 Discussion

4.4.1 Alpha diversity

Shannon (alpha) diversity was significantly higher in the sediment than in the water column, regardless of the water source type. As discussed in Chapter 3, the increased diversity in the sediment compared to the water column at freshwater sites is a result often reported in published scientific literature (Abia et al., 2018; Staley and Sadowsky, 2016).

Shannon diversity was highest in Autumn and lowest in Spring, which is a misalignment with the Summer-Winter dichotomy often reported in published scientific literature (Chen et al., 2018; Roberto et al., 2018; Shang et al., 2022; Shilei et al., 2020). Due to the sampling period being only one year (i.e., one iteration of each season), analysis of the influence of season on bacterial community composition in this study is somewhat limited as there is not data to assess whether this trend is recurrent or a divergence from an otherwise expected seasonal pattern.

4.4.2 Beta diversity and PERMANOVA

The NMDS of the Bray-Curtis distances of all the samples showed that the bacterial communities in the water and in the sediment are distinct. The water and sediment samples are visibly separated along the first axis and the PERMANOVA confirmed the centroids were significantly different. Similar results from ordination have been found in other freshwater studies where the sediment and water column were compared (Ibekwe et al., 2016; Nevers et al., 2020; Staley et al., 2016; Wang et al., 2018).

Comparing the PERMANOVA results for the water subset to those of the sediment subset reveals important similarities and differences in the extrinsic drivers of these bacterial communities. The significance of the physicochemical parameters remained the same between the water and sediment analysis except for pH, which was a significant explanatory variable for the water PERMANOVA only. This suggests that the bacterial communities from water and sediment are both affected by physicochemical water quality. Physicochemical

water quality parameters are essential to defining suitable habitats for microorganisms, therefore it was expected in this study that many parameters would be statistically significant. These results are mostly concordant with published literature reporting that water bacterial community composition is influenced by pH, alkalinity, and organic carbon content (Llirós et al., 2014), and that both water and sediment composition is affected by pH and electrical conductivity (Ibekwe et al., 2016). One study found that sediment community composition is significantly affected by localised pH (Liu et al., 2015), and this is potentially reflected in the current analysis.

Interestingly, discounted rainfall and temperature were significant in the water PERMANOVA but not in the sediment PERMANOVA. This suggests that the bacterial communities in the water are responsive to meteorological conditions, and sediment communities are not. The water column potentially provides a buffer to the sediment communities from fluctuations in seasonal conditions (Zhang et al., 2021). A recent study on freshwater lakes concluded that bacterial communities were dissimilar between the sediment and water because internal and external processes affect them differently, even if the bacteria originated from a common source (Nevers et al., 2020).

4.4.3 Connectedness and cohesion

While the extrinsic drivers of the bacterial communities were explored with the PERMANOVAs, the intrinsic dynamics were explored with connectedness and cohesion. The taxa with the highest positive connectedness were predominantly Proteobacteria, which was also the most abundant phylum. The taxa with the greatest negative connectedness were predominantly Actinobacteria and Proteobacteria, including two appearances each of the genus *Candidatus Planktophila* and the hgcl clade, common freshwater bacteria (Llirós et al., 2014). An ASV annotated as *Nitrospira* was reported as both highly positively and negatively connected to other taxa, suggesting that it was detected only within specific community compositions. By virtue of how cohesion is calculated, these highly connected taxa correspond to increased compositional stability in the sampled water sources. These taxa likely play important roles in the proper functioning of their environments (Dai et al., 2019). The cohesion metric does not describe the nature of taxon correlations (i.e., a positive correlation between taxa could be due to mutualism or response to a shared predator).

As connectedness is a measure of statistical correlation, the cohesion metric assumes that stronger cohesion (positive or negative) indicates a community with many taxa responding simultaneously to environmental fluctuations. Strong negative cohesion is a sign of community stability as it is related to lower compositional turnover (Herren and McMahon,

2017). Cohesion has been shown to be a strong indicator of Bray-Curtis dissimilarity (Danczak et al., 2018; Herren and McMahon, 2017), and this holds true in the current study, as shown by the PERMANOVAs. Cohesion was greater in water than sediment, and greater in river sites compared to on-farm storages, which is reflected in the NMDS plot of Bray-Curtis distances.

In future research, it would be interesting to see if a more marked difference in cohesion is visible in pristine versus disturbed sites; the on-farm and river sites in this study were in proximity to similar land uses and so no suggestion could be made whether community cohesion was linked to anthropogenic influences.

4.4.4 *E. coli* and community characteristics

According to the results of the correlation analysis in this study, *E. coli* is not strongly correlated to irrigation water pathogen potential (i.e., the IPPI), nor other bacterial community characteristics. In the water samples, only a weak positive Spearman correlation was reported between *E. coli* MPN and Shannon diversity. In the sediment samples, *E. coli* had a weak Spearman correlation to Shannon diversity and the IPPI. *E. coli* did not correlate to *Enterobacterales* abundance or cohesion measures. This is not an unexpected result, as the *E. coli* concentrations were derived from cultured bacteria and the other indices were metagenomically-derived. As discussed in Chapter 3, the metagenomic techniques used are not able to distinguish between dead bacterial cells and bacteria that were viable and active at the time of sampling. Therefore, a correlation between the two detection methods was never guaranteed.

4.4.5 Limitations

Many of the limitations discussed in Chapter 4 remain relevant in the current chapter. The water and sediment sampling, DNA extraction method, 16S rRNA sequencing, and the ASV filtering and analysis through the DADA2 pipeline likely introduced error and uncertainty at each step. Additional limitations to the methods used in this chapter relate to the statistical analyses.

The ordinations were performed on a subset of ASVs that were selected for their prevalence within the samples, good coverage in the SILVA taxonomic database, and membership to the five most abundant phyla. While subsetting was necessary to be able to ordinate and plot effectively, this analysis was ultimately performed on less than 1% of the taxa; this small subset was likely to be both common among samples and dominant within samples, which may have reduced the ability of the NMDS to capture the variation.

There has been some criticism or caution for using correlation-based approaches to study microbial communities (Carr et al., 2019) because biological dynamics and interaction structures are difficult to represent in numerical matrices. The authors of the cohesion metric (Herren and McMahon, 2017) emphasise that further work is needed to explore the implied meaning of positive and negative correlations in this context. I hypothesise that correlation-based techniques such as this are unlikely to be able to distinguish between biotic interactions such as predation, competition, and mutualism. Like the functional profile inference introduced in Chapter 4, the cohesion metrics are a relatively novel application within metagenomics and should be interpreted with care.

4.5 Conclusion

There was a strong correlation (~0.53) between cohesion and the IPPI in the water samples and a fair correlation (~0.36) for the sediment samples. This means that the potential for food safety pathogens increased as positive and negative cohesion became weaker. The implication of this is that stable bacterial communities appear to harbour fewer food safety risks. Community cohesion has been proven to decrease with increasing environmental stress (Hernandez et al., 2021), which may suggest that stressed environments are associated with increased human health repercussions.

5 Environmental, bloom-forming *Escherichia coli*

5.1 Introduction

Escherichia coli (*E. coli*) is a Gram-negative, rod-shaped bacterium in the Enterobacteriaceae family. Its primary habitat is the lower gastrointestinal tract of humans and other warm-blooded animals (Luo et al., 2011; Tenaillon et al., 2010). For this reason, *E. coli* is considered an indicator of faecal contamination in water bodies. However, there is substantial evidence that *E. coli* has become naturalised in the broader environment (Jang et al., 2017), meaning aquatic ecosystems with suitable nutrient availability and temperature ranges can support the survival and growth of *E. coli* outside of a host. The challenge presented to water managers is how to distinguish between *E. coli* strains that are accurate indicators of faecal contamination – especially disease-causing, pathogenic varieties – and environmental strains.

A phenomenon reported in Australian water reservoirs is environmental, "bloom-forming" *E. coli*. Under favourable conditions, environmental strains of *E. coli* can multiply to high densities, forming blooms. Concentrations of *E. coli* in these blooms extend beyond what can be explained by faecal contamination. For example, Lake Burley Griffin in the Australian Capital Territory (ACT) has historically recorded densities of over 10,000 *E. coli* colony forming units (CFU) per 100mL. This number is equivalent to a week's worth of sewage from over 470,000 people – higher than the local population – entering the lake (Power et al., 2005). The *E. coli* strains capable of such blooms do not carry genes that can cause disease in humans. However, their presence triggers the same operational response from water managers. If a water body is deemed unfit for use due to high *E. coli* counts, this could prevent access to recreational waters or disrupt the water supply chain.

A small subset of Australian water organisations has reported experiencing environmental *E. coli* blooms since they were first recognised sixty years ago. Yet, internal reports are hard to acquire. The body of published scientific literature on blooms in Australia is minimal (Mackay and Ridley, 1983; Nanayakkara et al., 2019; Power et al., 2005) and there is very little evidence of water quality events being recognised as environmental *E. coli* blooms in other countries (Solo-Gabriele et al., 2000).

Australian *E. coli* bloom events are rare overall but are recurrent in several reservoirs (Table 5.1). Blooms have appeared on the east coast of Australia from 1961 (Mackay and Ridley, 1983) and in a single event in a Western Australian reservoir in 2015. Blooms typically occur for three to four weeks, but extended monitoring has shown they can last up to three months (Sinclair, 2019). Increased nutrient levels in combination with warm water temperatures

(over 18°C) or rapidly falling dam levels (Bertone et al., 2019) can trigger bloom events. Nutrient influx events can include vegetation die-off, dust storms, algal blooms, and bushfires, but the exact nutrient responsible for the increased biosynthesis of these strains is currently unknown. One nutrient influx event – the September 2009 'Red Dawn' dust storm – triggered three separate *E. coli* blooms. The storm affected thousands of kilometres of the Australian east coast (De Deckker et al., 2014), causing blooms in the Warragamba (NSW), Grahamstown (NSW) and Hinze (QLD) dams (Bertone et al., 2019).

Recent modelling of a small sample of recorded blooms has shown that even under confirmed growth-promoting conditions, i.e., a preceding nutrient influx and decreasing dam volume, the probability of an *E. coli* bloom in a drinking water reservoir was only 44% (Bertone et al., 2019). Though *E. coli* blooms have been annual occurrences in the past, events appear less frequent in Australia after the Millennium Drought broke in 2010. Shifting climate patterns may increase bloom occurrence in the future (Bertone et al., 2019). For example, in Victoria, Pykes Creek Reservoir experienced inexplicably high *E. coli* counts in the summer and early autumn of 2018, and then again twice in early 2019 (Phelan, 2019). Experts indicated an undefined environmental bloom strain as the cause.

Table 5.1: Historical environmental Escherichia coli blooms in Australia, by location and year or approximate date (Bertone et al., 2019; Mackay and Ridley, 1983; Phelan, 2019; Power et al., 2005; Sinclair, 2019).

Location	Dates
Lake Burragorang and Warragamba Dam, NSW	Annually 1961-1967; 1974-1981 Four blooms 1984-1998 Two blooms 2008-2018
Lake Burley-Griffin, ACT	Annually 1975-2005
Hinze Dam, QLD	2009
Grahamstown Dam, NSW	December 2009 December 2010 March 2011
Maroon Dam, QLD	January 2014
Millstream Dam, WA	January 2015
Pykes Creek Reservoir, VIC	Late summer/early autumn 2018 January, February 2019

Blooms are not strongly associated with noticeable changes in physical water quality, such as turbidity (Bertone et al., 2019). The first sign of an impending bloom is typically an *E. coli* count 100-times higher than the previous reading, which can subsequently peak at around 10,000-fold (Sinclair, 2019). Therefore, routine water monitoring is essential for detecting these blooms. Still, bloom strains can be isolated from water even when *E. coli* counts are

low, i.e., non-bloom periods. Bloom-forming strains can persist in small numbers for prolonged periods between bloom events. One hypothesis is that the bloom-forming strains survive in the sediment longer than most other bacteria. The bacteria can then make use of the nutrient influx event when it occurs. *E. coli* are known to survive and multiply in sediment as it provides the necessary nutrients for growth and protects the cells from predators and environmental stressors such as ultraviolet radiation (Park et al., 2017).

E. coli lineages are believed to be widespread in the environment. Still, they are rarely identified because there is an understandably strong focus on research of commensal and pathogenic *E. coli* strains from human and animal sources. The *E. coli* genome is approximately 5Mb in size but may vary by more than 1Mb (Uyttendaele et al., 2015). This high degree of variability plays a role in the diversity of characteristics displayed by different *E. coli* strains, i.e., commensal, pathogenic, and environmental. Environmental *E. coli* have a relatively small genome (Touchon et al., 2020), which could explain how bloom strains can replicate so quickly and efficiently. Only a few strains of *E. coli* account for the majority of isolates in an environmental bloom. Whereas in faecal contamination, there is a high diversity of *E. coli,* reflecting the wide variety of gastrointestinal microbiota in a healthy individual (Tenaillon et al., 2010). For Australian reservoirs, blooms are dominated by strains from three *E. coli* phylogroups: A0, A1, and B1 (Power et al., 2005).

All environmental bloom strains isolated in Australia possess a Group 1 polysaccharide capsule originating from the genus *Klebsiella*. About 7% of all *E. coli* are encapsulated, but only bloom strains exhibit these Group 1 capsules (Nanayakkara, 2019). Group 1 capsules are indicated by the presence of the *galF* gene (Nanayakkara et al., 2019), encoded by the *cps* gene cluster. Encapsulated strains can be isolated from vertebrate hosts, soil, and water; bloom strains have been isolated from water only (Power et al., 2005). Among the bloom strains, several distinct Group 1 capsule types have been identified to date: KL16, KL31, KL53, KL49, KL60, KL63, and KL101. It is unlikely that encapsulation is the only genetic trait required for bloom status. Still, no additional genes reported are unique to bloom strains compared to other encapsulated *E. coli* strains.

Polysaccharide capsules are more likely to occur in free-living species than commensal species (Rendueles et al., 2017). Production of a capsule may confer several bloom strain characteristics: capsules can enhance survival and persistence by helping the cell to overcome predation, desiccation, osmotic stress, and ultraviolet radiation, and by facilitating ecological transitions and biofilm formation (Rendueles et al., 2017). Therefore, encapsulated strains may be more likely to persist in the environment than other *E. coli* strains and take advantage of increasing nutrients and ideal conditions to form blooms

(Sinclair, 2019). Additionally, bloom strains have a higher prevalence of genes that code for type 1 pilus synthesis, assembly and regulation (Nanayakkara et al., 2019). This type of pilus is required for *E. coli* to attach to abiotic surfaces. Further, the mucoid nature of bloom strain colonies reinforces the suggestion that they are well-adapted for growth in biofilms. However, there has been no reported isolation of bloom *E. coli* from a biofilm in the scientific literature.

Klebsiella pneumoniae and *E. coli* are both faecal coliforms, occur in the intestinal tract of mammals, and are opportunistic pathogens (Leclerc et al., 2001). *K. pneumoniae* is ubiquitous in the environment and frequently isolated from water supplies in the absence of faecal contamination. There is some evidence the *Klebsiella* capsule of *E. coli* is closely related to *K. pneumoniae* in particular, in terms of DNA sequence and structural arrangement (Rahn et al., 1999). Inheriting a *Klebsiella* capsule has potentially played a role in helping bloom *E. coli* to achieve a similar, free-living lifestyle to that of *K. pneumoniae*.

Water managers use *E. coli* to indicate the potential health risk from pathogenic microorganisms in water, introduced by faecal contamination. The consequence of environmental and bloom *E. coli* strains is that total *E. coli* counts are no longer proportional to faecal input as they have instead proliferated in the environment. The reliability of total *E. coli* as a measure of water safety can be questioned by the industry and detract from its utility. Fortunately, environmental *E. coli* strains and faecal *E. coli* strains respond similarly to water treatment and disinfection processes. This means the identification of bloom strains is a concern primarily for environmental water bodies and not for treated water stores.

Environmental strains of *E. coli* can present as atypical during normal laboratory processes, including enzyme-substrate analyses, growth at high temperatures, and biochemical tests. Bloom *E. coli* strains are said to have a mucoid colony appearance when cultured on agar plates. According to Power et al. (2005), the polysaccharide capsule is visible around individual bloom-forming *E. coli* cells when viewed under a microscope. Bloom strains can be identified by characterising their phenotypic and enzymatic properties. For example, *E. coli* bloom strain B1 diverges significantly in its biochemical profile from that of group A strains. This can cause them to be misidentified as *Citrobacter youngae* when using the API 20E system (Power et al., 2005).

A multiplex polymerase chain reaction (PCR) has been developed by Water Research Australia (WaterRA) and academics from the Australian National University to detect known bloom strains. The multiplex PCR detects *Klebsiella* capsule-positive strains, i.e., strains positive both for the *galF* gene and for previously characterised *E. coli* bloom strain capsule genes (Figure 5.1). This assay has been validated previously on pure cultures of

environmental isolates (Nanayakkara, 2019). The aim of this research chapter was to validate the assay at the Sydney Water laboratories and to perform the assay on samples collected from the upper Hawkesbury River region to identify the potential presence of bloom strains in fresh produce irrigation water sources.

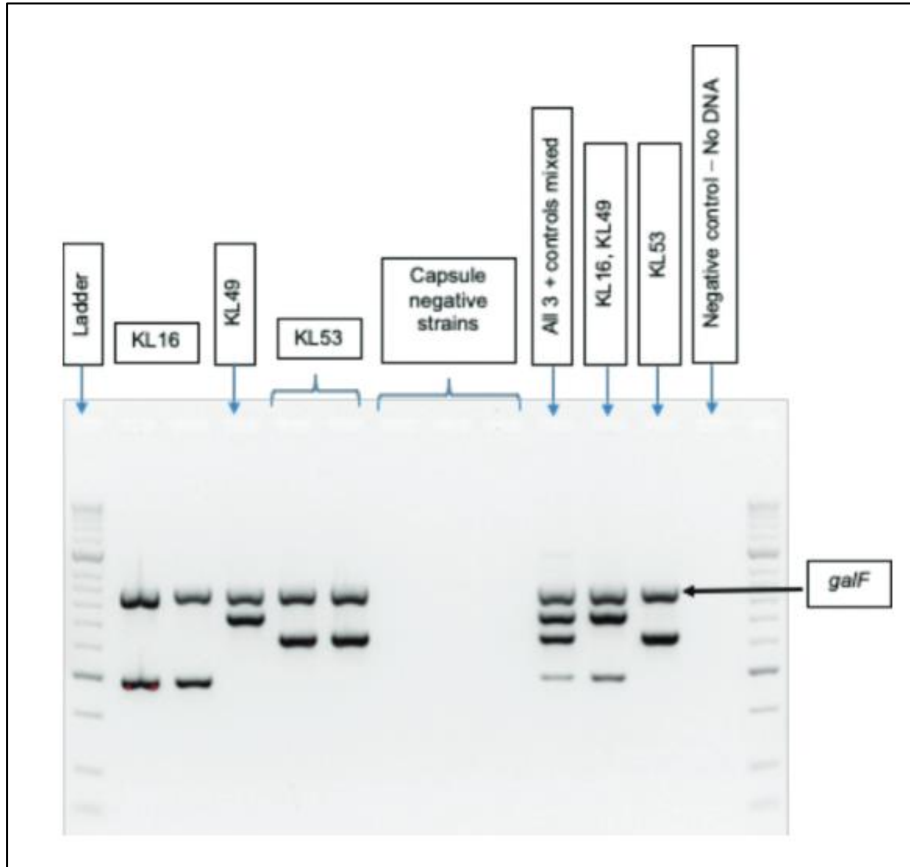


Figure 5.1: Gel electrophoresis of the capsule gene PCR products, individually and pooled. *galF* (668bp) indicates a Group 1 capsule is present.

5.2 Characterisation of a bloom isolate collection

5.2.1 Isolate information

The bloom-forming *Escherichia coli* isolates used in this project are detailed in Table 5.2. Isolates were generously provided by Professor David Gordon from the Australian National University, Canberra. *E. coli* isolates were preserved on cryobeads at -80°C to maintain the integrity of the cultures.

Table 5.2: Bloom-forming *Escherichia coli* isolate information, including identification, isolation location, phylogroup, and Group 1 capsule type

Isolate	Isolated from	Phylogroup	Capsule type
A2099.1.1	WA bloom	A1	KL101
A2099.1.2	WA bloom	A1	KL53
A2099.2.1	WA bloom	A1	KL63
A2106.2.1	WA bloom	A1	KL60
E267	East coast	B1	KL53
E704	East coast	A0	KL49
E2013	East coast	A1	KL16
E2061	East coast	A1	KL16
E8272	East coast	B1	KL53
E9432	East coast	B1	KL53
E9790	East coast	A0	KL49
E9791	East coast	A1	KL113
H288	East coast	A1	KL31

5.2.2 Growth and morphology characteristics

The *E. coli* bloom isolates were cultured on multiple agar types and their morphology was observed (Figure 5.2). The purpose of this was to confirm the mucoid colony appearance associated with the bloom strains (Mackay and Ridley, 1983; Power et al., 2005) and to record other aspects of their typical appearance on common agar types used in the Sydney Water microbiology labs. All isolates were incubated at 35°C overnight in a sterile nutrient solution before being streaked aseptically onto the agar.

5.2.2.1 MacConkey and tryptic soy agar (35°C)

The bloom isolates produced mucoid colonies on MacConkey agar, as expected (Figure 5.2a). These colonies appeared glossy or wet and had a glue-like consistency. Other bacteria such as *Klebsiella* and *Enterobacter* also produce mucoid colonies on MacConkey agar. This phenomenon is caused by the growth of a capsule, predominately formed from the lactose in the MacConkey medium. All the bloom isolates yielded off-white-coloured colonies on tryptic soy agar (TSA), which is typical of *Escherichia* species (Figure 5.2b). No other notable growth characteristics were observed for the bloom strains on TSA.

A. MacConkey agar

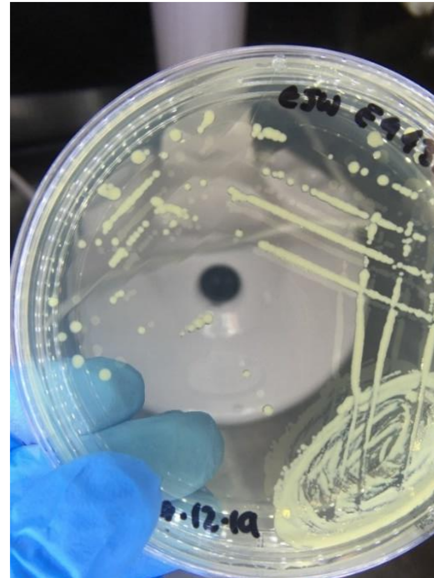
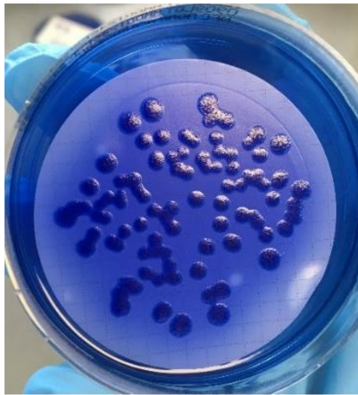

B. Tryptic soy agar (TSA)

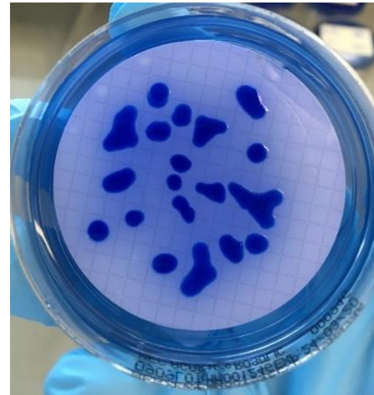
Figure 5.2: Example bloom isolates incubated on A) MacConkey and B) tryptic soy agar

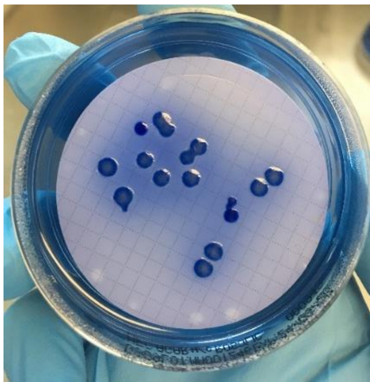
5.2.2.2 m-FC agar (35°C; 44.5°C)

Membrane faecal coliform (m-FC) agar is used for the detection of faecal coliforms by the membrane filtration technique. Faecal coliforms ferment lactose at higher temperatures (in this case, 44.5°C) and form blue colonies on the m-FC agar; other organisms produce grey colonies. The bloom *E. coli* isolates were grown on m-FC agar to assess two characteristics: if the bloom strains would be identified as faecal coliforms despite their environmental origin, and if growth of the B1 strains would be possible at 44.5°C.

Dilutions (10^{-9}) of overnight nutrient cultures were filtered through sterile 0.45 micrometre filter papers. The filters were placed on m-FC agar and incubated in metal canisters in water baths for 35°C for two hours, followed by 20 hours at 44.5°C. All isolates formed colonies using this method, despite expecting that B1 strains were not as thermotolerant (Power et al., 2005) and would not grow (Sinclair, 2019). However, the resuscitation of the cultures at 35°C for two hours in the water bath may have improved the growth of heat-stressed B1 strains, which has led to this discrepancy in experimental results.


A. Isolate E2013


B. Isolate E9432


C. Isolate A2099.2.1

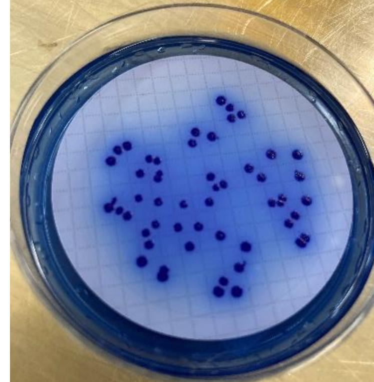

D. Lab strain ATCC 25922

Figure 5.3: Observed variation in *Escherichia coli* isolate morphologies as grown on m-FC agar, A and B are East coast isolates; C is a West Australian isolate, and D is a laboratory reference strain.

The growth on m-FC agar showed the most difference between isolate growth patterns. Some isolates grew large, flat, textured, dark-blue colonies (Figure 5.3a); some strains produced mucoid, irregular, bright-blue colonies (Figure 5.3b); and some strains had colonies with cream-coloured centres (Figure 5.3c). Grey- or cream-coloured colonies on m-FC agar suggest that the strains could not ferment lactose to acid and should be considered as non-faecal coliforms. The discolouration in the centre of a small number of bloom strain colonies is more likely to indicate an incomplete fermentation of lactose rather than an inability to do so since the colonies are ringed in blue. For reference, a typical *E. coli* morphology is shown in Figure 5.3d.

B1 isolates grew on the m-FC agar at 44.5°C, which was previously identified as too hot for growth (Sinclair, 2019). No published research is available to clarify the laboratory conditions under which this thermo-intolerance was previously observed. Therefore, we cannot confidently explain the conflict in reporting, but can suggest the resuscitation step (35°C for two hours) as a potential reason.

5.2.2.3 Tryptic soy agar (44.5°C)

To investigate thermotolerance on an additional agar type, all bloom isolates were struck onto TSA and incubated overnight at 44.5°C, the highest temperature incubator available. Again, all isolates were shown to grow at this temperature, including the B1 strains. However, some strains did appear to be more growth-limited relative to others, shown by the appearance of smaller, more scarce colonies.

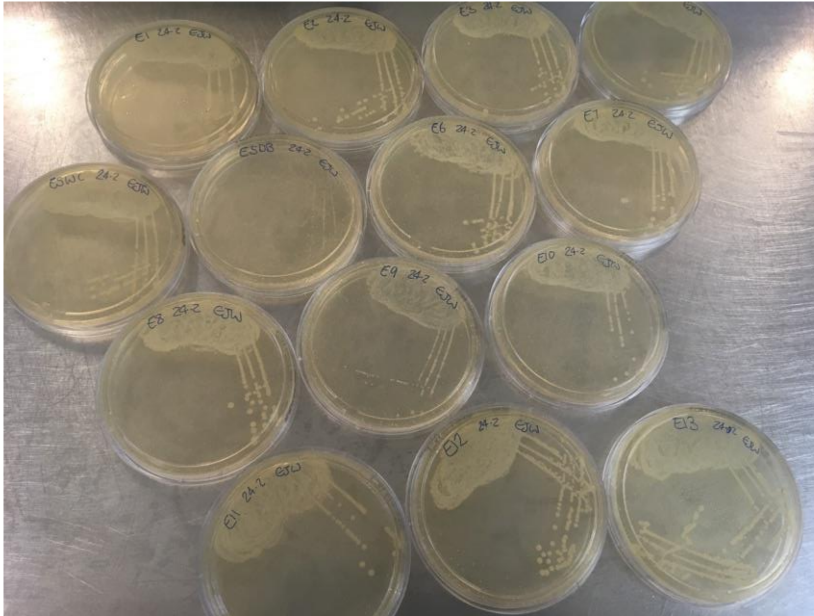


Figure 5.4: Growth of bloom isolates on tryptic soy agar (TSA) after incubation overnight at 44.5°C

5.2.3 Enzyme tests

The IDEXX™ Colilert substrate is designed to simultaneously detect total coliforms and *E. coli*. Two indicators, ortho-Nitrophenyl-β-galactoside (ONPG) and 4-methylumbelliferyl-β-D-glucuronide (MUG), are the major sources of carbon in the Colilert medium and can be metabolised by the coliform enzyme β-galactosidase and the *E. coli* enzyme β-glucuronidase, respectively. Metabolism of ONPG (ortho-nitrophenyl-β-galactoside) turns the substrate yellow; metabolism of MUG (4-methylumbelliferyl-beta-D-glucuronide) results in a blue fluorescence under long-wave ultraviolet light. While the Colilert medium is specially formulated to suppress the growth of non-*E. coli* organisms that can metabolise MUG, thus preventing false positives, it cannot detect the minority of *E. coli* strains that are MUG-negative.

To test each bloom isolate, 1ml of spiked nutrient solution was added to 99mL of sterile water, and this suspension was processed with the Colilert system according to the manufacturer's instructions. While all isolates were positive for total coliforms after incubation (i.e., the substrate was yellow in colour), not all isolates fluoresced under

ultraviolet light, which is the requirement for detecting and quantifying generic *E. coli*. Isolates E704 and E9790, which are both phylotype A0, were negative; all other isolates were positive.

This result initially caused confusion because the A0 phylotype had previously been reported as MUG-positive, and the phylotype B1 had been reported as MUG-negative (Sinclair, 2019). However, after repetition of the Colilert experiment to confirm accuracy, communication with the authors of the WaterRA assay confirmed that our results are accurate and have also been reported by Nanayakkara (2019). A1 and B1 phylotype strains are MUG-positive, and A0 phylotype strains are MUG-negative. If the Colilert™ substrate is the sole method of identifying bloom *E. coli*, the A0 strain will not be detected in water as it does not produce β-glucuronidase.

5.2.4 Biochemical profiles

The biochemical profiles of the isolates were investigated using the API 20E system (BioMerieux). API (Analytical Profile Index) 20E is a biochemical panel for identification and differentiation of members of the Enterobacteriaceae family of bacteria. It is a well-established method of manually identifying and differentiating Enterobacteriaceae to the species level based on the presence or absence of enzymatic activity. All positive and negative enzymatic test results are compiled to obtain a unique profile.

For each bloom isolate, a single colony was chosen from MacConkey agar and suspended in 9mL of sterile Ringer's solution. The solution was then processed through the API 20E system according to the manufacturer's instructions. The profiling was repeated for each isolate to account for sensitivities in individual assays.

All group A isolates were successfully identified as *E. coli* (Table 3). A0 isolates (5144552) were distinct from all others as they were positive for ornithine decarboxylase (ODC). The A1 isolates showed more variation in biochemical profile: 1044573, 5044552, 5044553, 5045542, and 5045552.

The B1 isolates were consistently assigned "low confidence" by the API 20E system. They ultimately could not be identified as *E. coli* by this method. These isolates (E267, E8272, E9432) were identified as most likely to be *Citrobacter youngae* with just 69.7% confidence. This result is in keeping with previous studies that also found *E. coli* isolates to be misidentified as *Citrobacter youngae* (Mackay and Ridley, 1983; Power et al., 2005).

5.2.5 Capsule typing

The capsule types of the isolates were confirmed using the polymerase chain reaction (PCR) assay as described in the WaterRA Project #1101 report, followed by visualisation with gel electrophoresis using standard protocols in the Sydney Water labs. The purpose of this was to confirm that the isolates provided were as described, and to confirm that ethidium bromide was not necessary to accurately visualise the results.

To extract DNA for the assay, overnight nutrient cultures of each isolate were transferred into sterile 50mL Falconer tubes and centrifuged at 5000 *g* for two minutes. The supernatants were discarded, and the pellet was transferred to sterile Eppendorf tubes. To lyse the bacterial cells, the Eppendorf tubes were heated to 95°C for 10 minutes in a thermocycler. After lysing, the Eppendorf tubes were centrifuged at 5000 *g* for 30 seconds to pellet the cell debris. The supernatants were transferred into clean Eppendorf tubes and the nucleic acid concentration was quantified on a Qubit platform.

As summarised in Table 5.3, the isolates were confirmed to possess their reported capsule types, thus validating the diagnostic assay. In addition, the diagnostic assay can detect capsule type KL113 (isolate E9791) as it is virtually identical to capsule type KL16 (Gordon, 2019, unpublished data).

5.2.6 Isolate characterisation summary

Table 5.3: Detailed summary of *Escherichia coli* bloom isolates, including identification, phylotype, Group 1 capsule type, primer set, expected amplicon size, colony growth at 44.5°C, MUG production, and API 20E biochemical profiles.

Isolate ID	Phylotype	Capsule type	Primer set	Amplicon size (base pairs)	Growth at 44.5°C	MUG	API 20E																					
							Code	Identity	ONPG	ADH	LDC	ODC	CIT	H2S	URE	TDA	IND	VP	GEL	GLU	MAN	INO	SOR	RHA	SAC	MEL	AMY	ARA
E704	A0	KL49	wcuI	542	+	-	5144552	Escherichia coli 1	+	-	+	+	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+
E9790	A0	KL49	wcuI	542	+	-	5144552	Escherichia coli 1	+	-	+	+	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+
A2099.1.1	A1	KL101	wcuN	558	+	+	5045542	Escherichia coli 1	+	-	+	-	-	-	-	-	+	+	-	+	+	-	+	-	-	+	-	+
A2099.1.2	A1	KL53	wcuE	437	+	+	1044573	Escherichia coli 1	+	-	-	-	-	-	-	-	+	+	-	+	+	-	+	+	+	+	+	+
A2099.2.1	A1	KL63	wscD	231	+	+	5045542	Escherichia coli 1	+	-	+	-	-	-	-	-	+	+	-	+	+	-	+	-	-	+	-	+
A2106.2.1	A1	KL60	wcqX	834	+	+	5045552	Escherichia coli 1	+	-	+	-	-	-	-	-	+	+	-	+	+	-	+	+	-	+	-	+
E2013	A1	KL16	wcsT	293	+	+	5044552	Escherichia coli 1	+	-	+	-	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+
E2061	A1	KL16	wcsT	293	+	+	5044552	Escherichia coli 1	+	-	+	-	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+
E9791	A1	KL113	wcsT	293	+	+	5044553	Escherichia coli 1	+	-	+	-	-	-	-	-	+	-	-	+	+	-	+	+	-	+	+	+
H288	A1	KL31	wctG	339	+	+	5044552	Escherichia coli 1	+	-	+	-	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+
E267	B1	KL53	wcuE	437	+	+	1004512	Citrobacter youngae	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	-	-	+
E8272	B1	KL53	wcuE	437	+	+	1004512	Citrobacter youngae	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	-	-	+
E9432	B1	KL53	wcuE	437	+	+	1004512	Citrobacter youngae	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	-	-	+
Escherichia coli ATCC 25922					+	+	5144552	Escherichia coli 1	+	-	+	+	-	-	-	-	+	-	-	+	+	-	+	+	-	+	-	+

+ = positive; - = negative. MUG = 4-methylumbelliferyl-β-D-glucuronide; ONPG = ortho-Nitrophenyl-β-galactoside; ADH = arginine dihydrolase; LDC = lysine decarboxylase; ODC = ornithine decarboxylase; CIT = citrate as sole carbon source; H2S = production of hydrogen sulphide; URE = urease; TDA = tryptophan deaminase; IND = indole test; VP = Voges-Proskauer test (acetoin); GEL = gelatinase; GLU = glucose*; MAN = mannose*; INO = inositol*; SOR = sorbitol*; RHA = rhamnose*; SAC = sucrose*; MEL = melibiose*; AMY = amygdaline*; ARA = arabinose*. (* = fermentation of)

5.3 Validation of the diagnostic assay at Sydney Water

This diagnostic assay has been designed to identify environmental 'bloom-forming' *Escherichia coli* and works off two criteria: 1) the tested strain is shown to possess a *Klebsiella* capsule gene, and 2) the strain is shown to possess a specific capsule type known to be associated with past blooms.

The procedure consists of two multiplex PCRs, nicknamed the "East Coast pool" and the "Western Australia pool". The primers are designed to detect a core gene that signals the presence of a *Klebsiella* capsule (*galF*, present in both pools), and a range of capsule types. The capsule types detected by each pool are relevant to strains previously identified in those geographical regions. The East Coast pool detects four capsule types (KL16, KL113, KL49, and KL53); the Western Australian pool detects five capsule types (KL31, KL53, KL60, KL63, and KL101). The PCR products are visualised with gel electrophoresis.

5.3.1 East Coast pool

The East Coast pool was successfully validated (Figure 5.5). The *galF* gene was visible at 668bp for each capsule type and the positive multiplex, and each capsule type produced a second band according to Table 5.4. An additional capsule type – KL113 – was amplified by the *wcsT* primer set, resulting in a product 293 base pairs long. KL113 is indistinguishable from KL16 through gel electrophoresis as their PCR products are the same length.

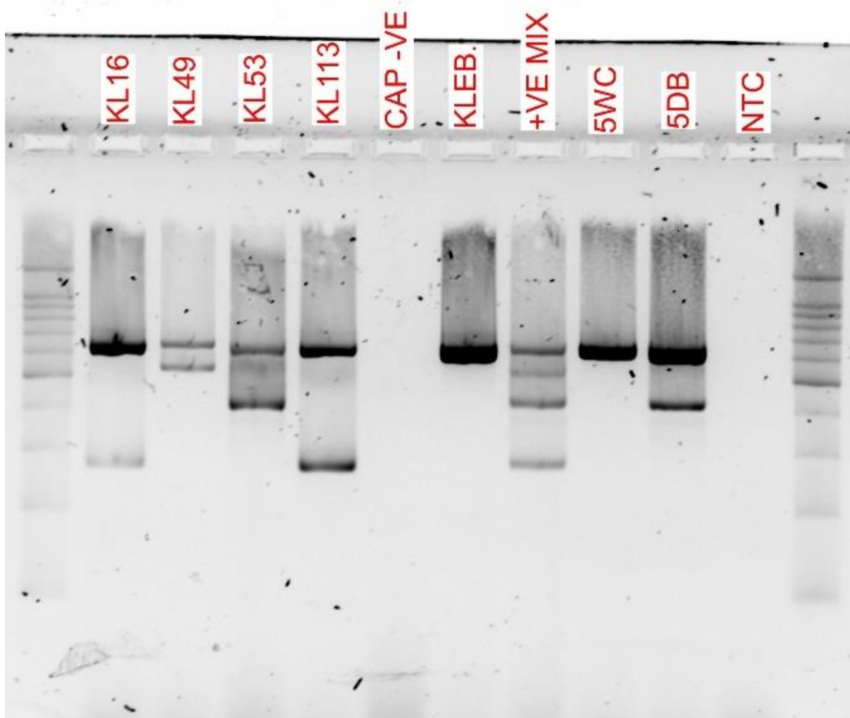


Figure 5.5: Electrophoresis gel showing amplicon size of the East Coast pool PCR products

5.3.2 West Coast pool

The West Coast pool was successfully validated (Figure 5.6). The *galF* gene was visible at 668bp for each capsule type and the positive multiplex, and each capsule type produced a second band according to Table 5.4.

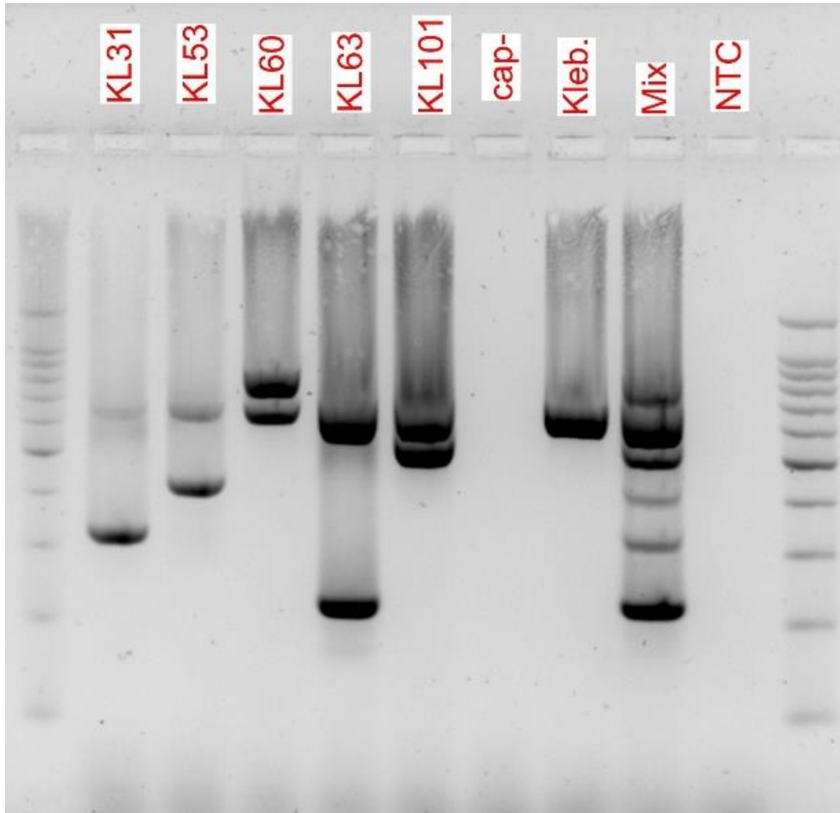


Figure 5.6: Electrophoresis gel showing amplicon size of the West Coast pool PCR products

5.3.3 Optimised protocol

The PCR was completed in 20ml reaction volumes. The concentrations and volumes of the primers and reagents in each reaction were as follows: 4ml of 5X MyTaq Red reaction buffer, 0.2ml of MyTaq HS DNA polymerase, 10mM of each primer, and 10-30ng of template DNA. The final volume was made up to 20ml with DNA-free water. For each assay, positive and negative controls were included: a combined DNA sample of target capsule types for each PCR pool, standardised to 10-30ng, and a non-template control, e.g., DNA-free water.

The thermocycler protocol was as follows: initialisation for 10 minutes at 94°C, 25 cycles of denaturation (94°C; 20s), annealing (57°C; 30s), and extension (68°C; 2m), followed by a final extension of 10 minutes at 72°C.

The PCR results were best visualised on a 2% agarose gel (0.01% GelRed) with 0.5X tris-borate-EDTA buffer (TBE) at 100V for 40 minutes. Approximately 6ml of a 100bp ladder (Promega, 1:17 dilution, stained with loading dye) and 6ml of each PCR product was

sufficient. Note: loading dye was not necessary for the PCR products as the MyTaq Red reaction buffer was already coloured. Successful amplification resulted in all products as described in Table 5.4.

Table 5.4: Details of the primer sets used in the diagnostic assay, including target genes, and PCR product sizes

Primer set	Target gene	Product size (bp)
galF	galF	668
wcuI	KL49	542
wcuE	KL53	437
wcsT	KL16/KL113	293
wcqX	KL60	834
wcuN	KL101	558
wctG	KL31	339
wcsD	KL63	231

5.4 Application on pooled environmental samples

5.4.1 Spiked environmental samples

The diagnostic assay had previously only been applied to pure cultures; the assay had not yet been trialled on pooled environmental DNA (eDNA). The benefit of an assay applicable to eDNA is that the potential detection of bloom strains in water samples would be quicker; there would not be a need to first isolate *E. coli* strains and to make overnight cultures. The main disadvantage of applying this PCR to eDNA is that *Klebsiella*, abundant in the natural environment, also possess the *galF* gene (i.e. the common capsule gene) and could lead to false positive detections of bloom strains.

To do this, 100mL of water sampled from Sydney Water site AN4 was spiked with one colony of each capsule type. The spiked water was filtered through a sterile, 0.45-micrometre filter paper. The filter paper was processed using a PowerSoil Pro Kit (Qiagen) for DNA extraction, with 2 x 30s at 8m/s on a bead beater. This process was repeated to produce duplicate samples (sample "A" and "B"). Additionally, 100mL of water from the same sample that had not been spiked was also filtered, providing a purely environmental sample ("Env."). A PCR was performed with both the East coast primer pool and the Western Australian primer pool, including mixed positive controls and non-template controls for each. The results were visualised using electrophoresis gel (Figure 5.7).

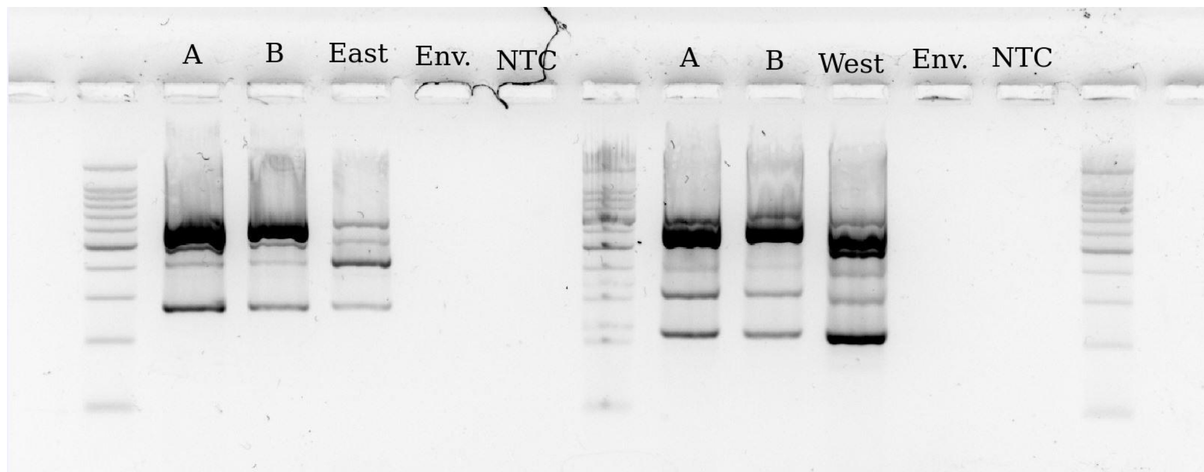


Figure 5.7: Electrophoresis gel of the spiked samples A and B, and the East Coast and West Coast pool PCR products

The results show that the environmental water sample ("Env.") and the non-template control (NTC) were negative for the *galF* and capsule-type genes. We can conclude that the PCR was free of contamination and that the environmental sample was free of *Klebsiella.* Therefore, the amplified bands visible in samples A and B are due to the spiking of the water sample and represent one bloom-forming colony per 100mL of water. All capsule types were visible on the gel, albeit difficult to distinguish at 600bp amplicon length where the *galF* product (668bp) is overrepresented.

5.4.2 Screening of the upper Hawkesbury River

Water and sediment samples were taken from the upper Hawkesbury River region monthly from September 2018 to August 2019 as part of a study on bacterial community composition. Environmental DNA (eDNA) extracted from these samples was used to screen the area for evidence of *E. coli* bloom strains. There had been no previously detected bloom event in this part of the Hawkesbury-Nepean River system, but water quality conditions had been ideal for bloom formation during part of this period: decreasing dam volume, elevated water temperatures, and a bushfire season. Lake Burragorang and Warragamba Dam, which feed into this system, historically experience blooms (Table 5.1), and so it is possible that *E. coli* bloom strains are also present further through the catchment. The sediment samples were investigated with the water samples because sediments provide favourable conditions for faecal indicator bacteria (Haller et al., 2009); it is possible that bloom strains were persisting in the sediment in the absence of optimal bloom-forming conditions. However, despite testing water column and sediment eDNA extracts with both elevated and low *E. coli* counts, no evidence of bloom strains could be found. More work is needed to determine the reason for this.

5.5 Discussion and conclusion

The presence of bloom strains, though not detected in environmental samples from the upper Hawkesbury River in this study, challenge the usefulness of *E. coli* as an indicator of recent faecal contamination. Bloom strains can potentially mask the presence of more dangerous, pathogenic strains in water. In reservoirs and other water bodies where these bloom strains are known to be detected, caution is needed if using *E. coli* as an indicator of microbial water quality.

The polymerase chain reaction (PCR) detection assay described in the WaterRA Final Report has been successfully validated at the Sydney Water labs. Further, the assay has been optimised at lower primer concentrations, which is more economical. It should be noted that the capsule type KL113 is indistinguishable from capsule type KL16 and will show a band at 293bp in an electrophoresis gel. This is unlikely to be problematic as it is still a successful detection of an *E. coli* bloom strain. However, it is possible to use the assay to screen pooled DNA from environmental water samples, thus bypassing the need to isolate bloom strains and culture them overnight. The results of a PCR on such samples should be considered with caution as *Klebsiella* may cause a false positive detection.

5.5.1 Future research

The current assay developed by researchers at ANU during the WaterRA project is a multiplex PCR with agarose gel electrophoresis. The primers and thermocycling protocol are designed for use on pure isolates, giving a presence-absence result for the environmentally blooming *E. coli* capsule types. As reported above, the assay can also be used on pooled eDNA, with or without the pre-enrichment step of overnight incubation. It is important to note that bloom strains are often present during non-bloom periods; confirming that one or more bloom strains are present does not guarantee that the strain responsible for the elevated *E. coli* counts has been identified. An extension to this research is developing a quantitative (or real-time) PCR to identify which strain is most abundant and therefore the cause of the bloom. There are several approaches that can be taken when developing a qPCR.

Dye-based and probe-based qPCR assays are used to detect and quantify DNA. Dye-based qPCR uses a fluorescent dye that binds to DNA and generates fluorescence that is proportional to the amount of target DNA. Probe-based qPCR uses a fluorescently labelled probe that bonds specifically to the target DNA, generating fluorescence when cleaved by the polymerase. The choice of method depends on the specific use and experimental design.

5.5.1.1 Dye-based assay

The addition of an intercalating dye would make the current assay quantitative. The resulting assay would be a series of single-plex reactions. This could be performed on the Rotor-Gene, which is faster than running a gel. The protocol would likely need to be optimised to work with the dye. A dye-based assay is less expensive than a probe-based assay, but potentially less accurate. A few studies have successfully multiplexed dye-based assays by developing primers with distinctive melting temperatures. If performing a melt-curve analysis, you could distinguish between the genes present by identifying peaks at different temperatures.

5.5.1.2 Probe-based assay

Probe-based assays have increased sensitivity and specificity compared to many dye-based qPCRs. The Rotor-Gene has five channels, meaning an assay can be developed that detects up to five targets, or four targets plus an internal control. Two assays would be needed to cover all seven identified capsule types. Purchasing probes would increase the cost of the assay.

The current primers were not designed with a probe-based assay in mind. This means that the PCR product size for several of the current primer sets is larger than the ideal range (amplicon length greater than 150 base pairs), meaning the efficiency of the reaction may be too low to be reliable (Ditommaso et al., 2015). This might not be an issue if using pure isolates but could be limiting if using this assay for pooled environmental DNA. The primers could be redesigned along with the design of the probes. Redesigning the primers would have several benefits: a chance of a smaller PCR product size, which is likely to increase the efficiency of the PCR, and the primer redesign may make the assay more applicable to environmental samples.

5.5.1.3 Metagenomics

Beyond PCR tools, it may be possible to develop metagenomic-based assays for the identification of bloom strains. The *galF* gene is diagnostic because only bloom strains of *E. coli* exhibit the Group 1 capsules (Nanayakkara, 2019), and therefore could be used as a marker gene. However, the gene is 668 base pairs long, which is too long for short read length platforms such as the Illumina MiSeq used in Chapters 3 and 4. In recent years, there have been advances in sequencing technologies that can generate longer read lengths, such as PacBio and Oxford Nanopore sequencing. These technologies have the potential to identify larger genes and genomic regions that may be difficult to identify using Illumina sequencing alone.

6 Discussion

6.1 What was the extent of microbial contamination in the studied irrigation water sources?

Water used for fresh produce irrigation in Australia is not held to the same standard as drinking water, and therefore we cannot expect it to be of the same high quality. Specifically, water designated for extraction by irrigators is not managed as a drinking water supply. This means that it is generally not monitored for its microbial quality and is not covered by health incident protocols. This has the potential to be detrimental to the fresh produce industry by putting at risk the safety and quality of the food produced. However, there is a waterborne and foodborne disease burden already present in Australia, and the reputation of the fresh produce industry is still considered excellent.

One aspiration of this thesis was to ascertain the extent of microbial contamination in Australian irrigation water sources. While fresh produce growers test water sources in keeping with their responsibilities under certification schemes, this data has never been compiled. As such, we do not know the typical microbial profile of Australian irrigation water sources. A field sampling program was undertaken to generate sufficient data for a preliminary analysis. The upper Hawkesbury River study only comprised of 96 water and 94 sediment samples in a single year from a single growing region, but it did reveal important insights on what might be the broader microbial risks associated with irrigation water sources in Australia.

Every 100mL water sample collected during the field study was positive for generic *E. coli.* Beyond the confluence of the Nepean and Grose Rivers, all the water sampling sites had a mean *E. coli* concentration close to or above the Freshcare certification limit of 100 cfu per 100mL. Applying the diagnostic assay for bloom-forming strains of *E. coli* did not reveal any positive samples for the *galF* gene, and so these concentrations cannot be discounted as environmental strains. The provision of data from WaterNSW allowed for an analysis of 1,852 datapoints from a wider area, and the results reiterated that surface water sources are generally contaminated with bacteria to some extent.

The taxonomic profiles of the water and sediment communities revealed that a considerable number of potentially pathogenic genera were present in the samples, including *Aeromonas, Bacillus, Campylobacter, Clostridium, Enterobacter, Enterococcus, Escherichia/Shigella, Listeria, Pseudomonas, Staphylococcus,* and *Yersinia*. Specifically, the human pathogens *B. cereus*, *C. coli*, *L. monocytogenes*, *S. aureus*, *E. faecium,* and *E. coli* were identified. The

application of NGS techniques to identify these pathogens does not distinguish between viable and nonviable bacterial cells, and the presence of virulence strains was not confirmed, but this does suggest that the sampled irrigation water sources contain a variety of human pathogens.

6.1.1 How did the sediment contribute to the microbial loads?

The importance of sediment as a source of bacterial contamination in water is mostly reinforced by research within this thesis. The sediment was an important reservoir of *E. coli,* as demonstrated in Chapter 2, and of potentially pathogenic taxa, as demonstrated in Chapter 3. The beta diversity analysis in Chapter 4 showed that the water and sediment bacterial communities were significantly different, which suggests that resuspension from the sediment to the water column was not extensive at a whole community level.

6.1.2 Which environmental and physicochemical water quality parameters were related to increased contamination?

This study allowed for a comprehensive list of environmental and physicochemical parameters to be investigated for their correlation with increased microbial contamination. Environmental and physicochemical parameters were included in the generalised additive model (GAM) of the Hawkesbury field sampling dataset, the regression and classification random forest models of the Greater Sydney dataset, and the beta diversity analyses.

Antecedent rainfall stood out as the most influential parameter for microbial water contamination, as reflected in other studies (Falardeau et al., 2017; Rock et al., 2016; Wilkes et al., 2009). Rainfall was a significant ($P < 0.05$) predictor of *E. coli* in the Hawkesbury GAM and contributed substantially to the performance of the regression and classification random forest models. Additionally, rainfall was a significant driver of beta diversity in the water samples.

6.2 Are the studied water sources suitable for the irrigation of fresh produce?

The Freshcare certification limit for the microbial quality of pre-harvest water is 100 *E. coli* cfu per 100mL. In Chapter 2 (page 21), *E. coli* concentrations from two datasets were analysed. In the Hawkesbury dataset, *E. coli* concentrations (n = 96) ranged from 3.1 to >2419.6 MPN per 100mL, with a Kaplan-Meier mean value of 199 MPN $100mL^{-1}$. A total of 35 samples (36.5%) were greater than 100 MPN $100mL^{-1}$. Six of the eight sampled sites had Kaplan-Meier mean *E. coli* values greater than 100 MPN $100mL^{-1}$ over the 12-month period of sampling. In the Greater Sydney dataset, *E. coli* concentrations (n = 1,852) ranged from

<1 to 26,000 organisms 100mL^{-1}, with a mean value of 134.9 organisms 100mL^{-1}. A total of 316 samples (17.1%) were greater than 100 organisms 100mL^{-1}. Using the Freshcare certification requirement as the definition of “safe”, and considering the above summary statistics, these irrigation water sources were not suitable for direct use on fresh produce a substantial number of times.

Many other certification schemes present in Australia do not have a prescriptive *E. coli* limit (Table 1.2), for example, SQF and HARPS. Under these schemes, the suitability of the irrigation water source would have been at the grower’s discretion, with the aid of a risk assessment. GLOBAL.G.A.P. allows for its growers to fall back on WHO guidelines for wastewater reuse where legislation doesn’t exist locally, meaning the limit is 1,000 cfu per 100mL. In this case, most but not all water samples would have given acceptable microbial results for use on fresh produce.

An important aspect of the current Freshcare standard is that if growers use a pathogen reduction step, the water source may potentially not need to meet <100 cfu *E. coli* per 100mL. The standard describes a pathogen reduction step as a “process which results in at least a 2-log reduction in the number of viable pathogens on a product or in water” (Freshcare Ltd, 2020). Using the highest *E. coli* concentration from the Greater Sydney dataset as an extreme example, this would mean that a water contamination level of 26,000 cfu per 100mL could theoretically be reduced to 260 cfu before application on a crop and still comply with certification.

It is difficult to determine whether a food safety outbreak would arise from the use of the studied irrigation water sources, even with the molecular identification of several known outbreak agents such as *E. coli* and *L. monocytogenes*. There is no data on the irrigation method, the timing of irrigation prior to harvest, or use of pathogen reduction steps in the Hawkesbury growing area, all of which would influence the persistence of pathogens on the crop. Therefore, this thesis is not suggesting that actual human illness has occurred as a consequence of growers using these water sources.

6.3 What additional information did next-generation sequencing contribute?

This thesis produced metagenomic 16S rRNA data for communities of bacteria that were likely to be largely unculturable. An enormous number of unique amplicon sequence variants (ASVs) were able to be generated by the DADA2 workflow, from which detailed taxonomic profiles, inferred functional profiles, and measures of community diversity and cohesion could be produced. The identification of potentially pathogenic genera and species was the

most notable contribution from NGS in this study because it is immediately informative to the fresh produce industry.

Taxonomic databases are comprehensive and still expanding but cannot be considered well-curated for food safety applications yet. There is no standard method for applying NGS to fresh produce food safety, which makes it difficult to compare across studies and prevents the compilation of data industry-wide. There is also no proposed level of harm associated with NGS data, i.e., the level of abundance at which the detection of genetic material equates to actual risk of human illness.

NGS is useful to the fresh produce industry and associated research groups as a means of increasing our understanding of microbial quality and safety, but in my opinion, it is unlikely to become standard practice to screen for pathogens in the near future. As it stands, there are too many inherent biases in the selection and implementation of NGS methods, and the results need to be critically interpreted to be useful. Most of all, due to the limited success in annotating genetic sequences to the species level using this approach, e.g., 0.61% in this study, the probability of false negatives is overwhelming. The fresh produce industry needs conclusive scientific evidence on which to base its requirements for growers. For this reason, NGS is unlikely to become routine.

Longer sequencing read lengths, i.e., greater than the 2x300bp chemistry currently offered by the Illumina MiSeq platform, has the potential to improve the metagenomic detection of species in irrigation water sources. Assembling longer contigs would improve the accuracy of the species assignment by providing more complete genetic information with which to resolve ambiguities. However, other molecular approaches may be a better fit. Whole genome sequencing (WGS) is already employed for traceback investigations during and following foodborne disease outbreaks and has been successful at identifying the root cause, leading to improved industry practices. In terms of a more proactive application of molecular science in food safety, microbial source tracking (MST) is gaining popularity in the drinking water industry. The MST approach is more targeted than community-based monitoring and can be highly sensitive and specific to human pathogens.

A combination of metagenomics and MST concepts may also be a useful application of environmental DNA concentrated from water sources. Molecular markers for animals such as cattle/sheep, waterfowl, and domestic animals (i.e., non-microbial, non-pathogenic) can be used to identify the source of water contamination (Warish et al., 2015). The underlying assumption is that the hydrologic processes that caused animal markers to be present in a water sample also facilitated the transport of zoonotic pathogens to the water source, thus informing water source risk assessment (Devane et al., 2018).

6.4 Was *E. coli* a useful indication of microbial risk?

Using *E. coli* as an indicator infers that pathogens may be present because it is assumed to originate from the same sources as pathogens, i.e., faecal contamination. For this reason, it was hypothesised that the concentration of *E. coli* as enumerated with IDEXX Colilert would be strongly correlated to the irrigation pathogen potential index (IPPI), which was derived from genera associated with enteric illness. *E. coli* and the IPPI had a rank correlation of 0.134 in the water and 0.232 in the sediment, which are considered weak links.

Grab samples for *E. coli* cannot be assumed to represent comprehensive water quality or necessarily assure microbial safety. The shortcomings of *E. coli* as an indicator have been outlined extensively in the scientific literature (Mendes Silva and Domingues, 2015; Wen et al., 2020). The default assumption that *E. coli* is entirely due to human or animal faecal contamination is overly precautionary due to the existence of environmental strains (Nanayakkara, 2019; Nanayakkara et al., 2019), but researchers are unlikely to get sufficient data to provide the industry with a defendable and widely applicable lower proportion. Conversely, there are both pathogenic (Brooks et al., 2008) and environmental (Power et al., 2005) strains of *E. coli* that are not enumerated in commonly used enzyme substrate methods such as IDEXX Colilert, and therefore many *E. coli* measurements are also not accurately precautionary. However, without a suitable alternative validated for use in Australian water sources and readily available for implementation, the fresh produce industry should retain generic *E. coli* as its preferred indicator.

Another area of future research that could be considered, is to extend the current project to the metagenomic analysis of viral and protozoan communities in water sources. Rusiñol et al. (2020) described the virome, bacteriome and parasitome of different irrigation water sources to evaluate the associated health hazards more comprehensively. A similar strategy could be applied to Australian water sources to identify viral and protozoan pathogens and assess their correlation to generic *E. coli*, or to extend or augment the IPPI.

6.5 Recommendations

6.5.1 Microbial limits and risk assessment

Hard borders for water quality are debatable because they are not always realistic. It can be helpful to define “safe” and “unsafe” using a defined threshold for many reasons, such as ease of use and clear direction on when action should be taken. Yet, it is also difficult to defend why 99 cfu *E. coli* per 100mL is acceptable but 100 cfu is not, to use the Freshcare irrigation water limit as an example. It may be for this reason that some other certification

schemes do not prescribe a hard border. This challenge is reflected in the recent changes to the proposed FSMA Rule on Agricultural Water in the USA (FDA, 2021). If adopted, the requirements in this proposed rule would substitute systems-based pre-harvest agricultural water evaluations for the previous Produce Safety Rule's pre-harvest microbiological quality criteria and testing requirements. To reduce the risks associated with pre-harvest agricultural water, these assessments would be used to identify conditions that are reasonably likely to introduce known or reasonably foreseeable hazards into or onto produce or food contact surfaces. They would also be used to determine whether corrective or mitigation measures are required.

Most irrigation water quality guidelines or requirements are empirical, fixed standards focusing on microbial indicator organisms, but risk assessment- or management-based approaches can offer more flexibility in specific situations (De Keuckelaere et al., 2015). Currently, the Freshcare assessment for pre-harvest water (Freshcare Ltd, 2020) has a risk matrix with two outcomes: high risk, and low risk. The risk assessment is based on the method and timing of irrigation, the type of produce grown, and the use of a pathogen reduction step. The risk assessment does not consider the water source (except to exclude its use when > 100 *E. coli* cfu per 100mL in high-risk scenarios) or microbial hazards that would lead to the contamination of the water supply. A shift from fixed numeric standards to a risk-based framework may be feasible for Australian fresh produce production.

The Australian Drinking Water Guidelines (ADWG) advocates for a risk-based framework (NHMRC and NRMMC, 2011). In the ADWG, the microbial quality of water in a drinking water catchment is managed to meet a health-based target (HBT). Briefly, a sanitary survey and microbial indicator assessment was performed to quantify the level of risk posed by hazards in the catchment. The level of risk is a continuum rather than an assessment of “safe” or “unsafe”. An assessment is also performed for the water treatment plants to ensure that the disinfection processes can overcome this risk. If so, the HBT is met. A similar framework could be pragmatic for fresh produce growers where their implemented pathogen reduction steps are adjusted to meet the microbial challenge posed by their water source, within reason.

6.5.2 Monitoring frequency and data collation

Microbial water quality monitoring is time-consuming and expensive, but that effort is justified to better define the health risks associated with irrigation water. Routine water sampling could reduce uncertainty. The fresh produce industry is aware of this because Freshcare certification requires monthly sampling during water source use. However, currently, four consecutive tests below 100 cfu *E. coli* are enough to classify the water source as

acceptable, followed by an annual test to confirm. In contrast, the Red Tractor assurance scheme in the UK denotes different risk levels and required test frequencies for different water sources, e.g., town water has a low risk level and should be tested annually, but canal water has a high risk and should be tested monthly. Interestingly, research has show that in a field, sampling a larger number of randomly-located samples is better than typical sampling strategies at detecting hazards (Quintanilla Portillo et al., 2022). It may be useful to investigate whether randomisation is also beneficial in sampling on-farm water storages.

There would be enormous benefit for the Australian fresh produce industry in exploring whether microbial water quality data collected by individual growers could be compiled into a single database. Freshcare is Australia's largest certification scheme for fresh produce food safety and may be in a position to do so. A microbial water quality database would be valuable to the development of predictive models such as those in Chapter 2 and in the AgWater App. Ideally, growers who extract irrigation water from the same water body, as many growers in the Hawkesbury region may do, could benefit from shared knowledge about the condition of their water supply.

6.6 Conclusion

This thesis furthers current knowledge of the microbial and physicochemical quality of Australian irrigation water sources by:

- Generating a comprehensive water quality dataset focusing on water sources potentially used for irrigation in the Greater Sydney area of New South Wales,
- Exploring the statistical relationships between environmental and water quality parameters and concentrations of *E. coli* in these surface water sources and leveraging them to build a preliminary predictive model,
- Outlining the taxonomic and functional profiles of the bacterial communities in the water column and sediment,
- Using novel techniques to explore trends in risk, such as the development of the irrigation pathogen potential index (IPPI) and the application of the cohesion metric, and
- Validating a diagnostic assay for the identification of bloom *E. coli* strains in Australian waters.

Overall, this thesis contributes to the ongoing discourse on the challenges associated with assigning risk to irrigation water sources and provides food for thought on the future application of metagenomics in fresh produce food safety.

7 Supplementary material

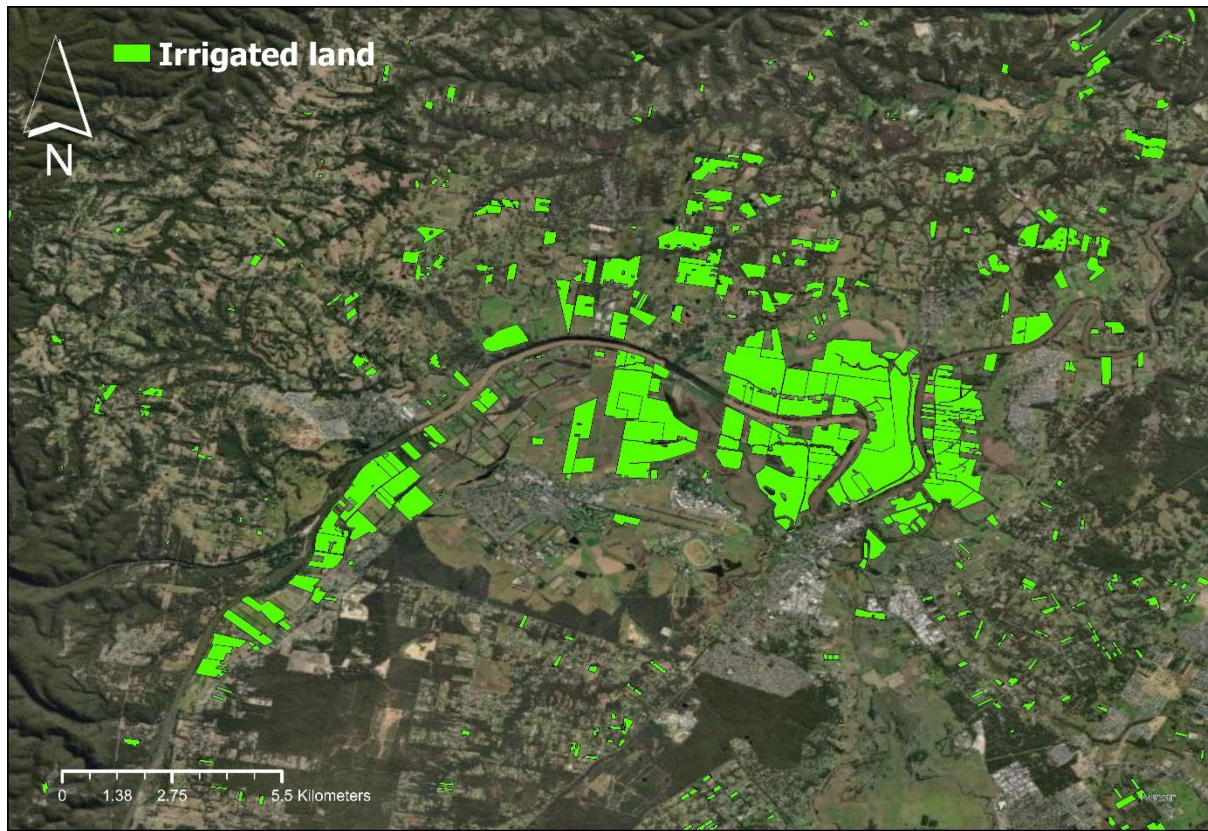


Figure 7.1: A map of the upper Hawkesbury River region, with irrigated land for seasonal or perennial horticulture or other crops indicated (“NSW Landuse 2017 v1.2,” 2020)

Table 7.1: Particle size analysis data for the sediments of the Hawkesbury field sampling sites, calculated according to the soils texture grid. All hydrometer readings corrected with a blank solution. 4:48 minute reading based off temperature of 20°C.

Site	Sediment collected (g)	Gravel weight (g)	Oven dry weight (g)	4:48 minute read	8:00 hour read		difference		Oven dry sand weight (g)		Total	Texture
				silt + clay (g)	clay (g)	clay (%)	silt (g)	silt (%)		sand (%)	(g)	
Bushells Lagoon	315	0	30	25	12	40.0	13	43.3	5.3	17.7	30.3	Clay/clay loam
Grose River	597	86	100	3	2	2.0	1	1.0	97.8	97.8	100.8	Sand
Hanna Park	343	10	40	4	2	5.0	2	5.0	36.8	92.0	40.8	Sand
Inalls Lane	439	0	30	9	6	20.0	3	10.0	20.5	68.3	29.5	Sandy loam/sandy clay loam
Nepean River	423	32	100	2	1	1.0	1	1.0	99.4	99.4	101.4	Sand
Punt Road	442	4	100	4	3	3.0	1	1.0	95.4	95.4	99.4	Sand
South Creek	526	8	30	4	2	6.7	2	6.7	26.4	88.0	30.4	Sand
Yarramundi Lagoon	338	0	30	20	9	30.0	11	36.7	11.5	38.3	31.5	Clay loam

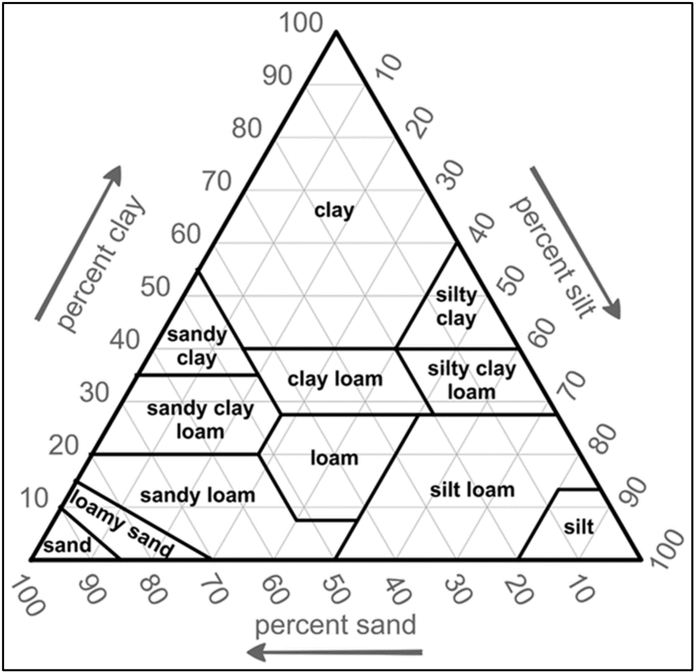


Figure 7.2: Soil texture triangle to determine textures given the sand, silt, and clay percentages of a sample

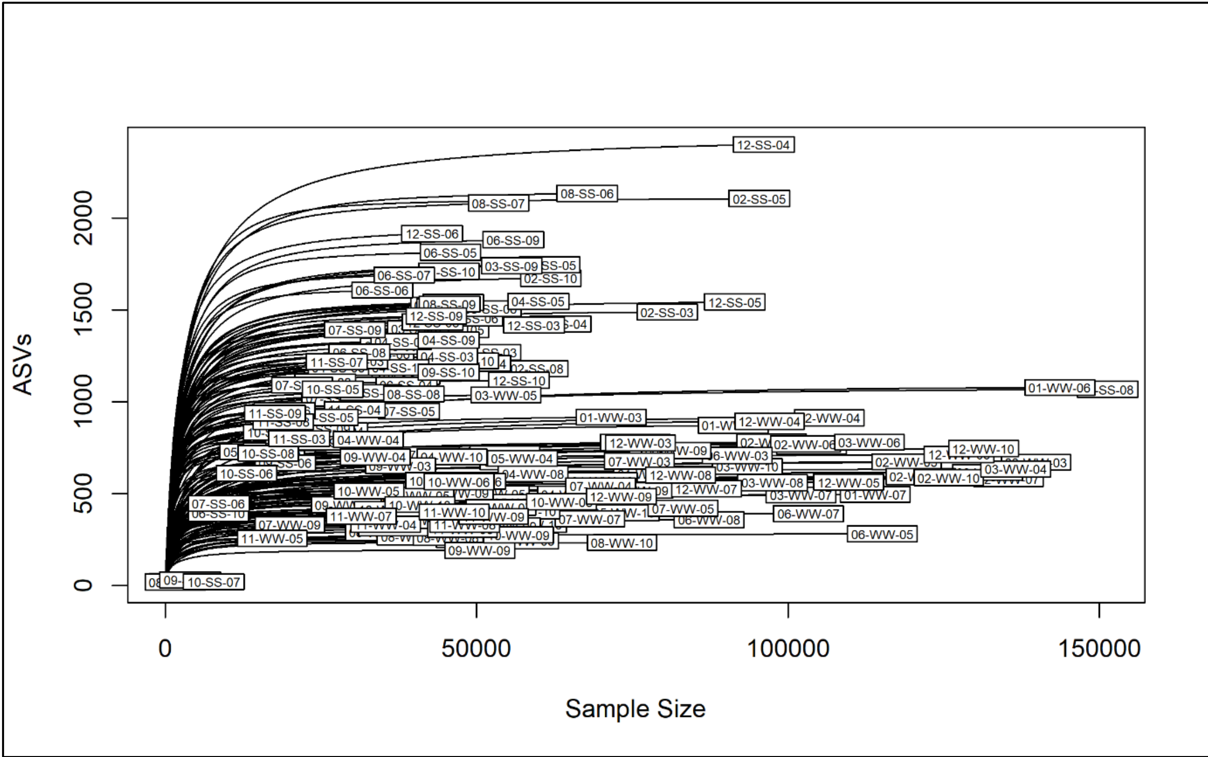


Figure 7.3: Rarefaction curves for the water and sediment samples (n = 190). Appropriate sampling depth is achieved when the rarefaction curve flattens, i.e., no further unique ASVs are generated from sampling more sequences.

Table 7.2: Successful species assignments for the genera included in the irrigation pathogen potential index (IPPI)

Genus	Species
Aeromonas	caviae
Aeromonas	media
Aeromonas	molluscorum
Aeromonas	popoffii
Aeromonas	sobria
Aeromonas	veronii
Bacillus	asahii
Bacillus	cereus
Bacillus	drentensis
Bacillus	murimartini
Bacillus	pseudofirmus
Bacillus	simplex
Campylobacter	coli
Enterobacter	asburiae
Enterobacter	cloacae
Enterobacter	muelleri
Enterococcus	cecorum
Enterococcus	faecium
Enterococcus	rivorum
Escherichia/Shigella	coli
Listeria	monocytogenes
Pseudomonas	alcaligenes
Pseudomonas	argentinensis
Pseudomonas	chlororaphis
Pseudomonas	cichorii
Pseudomonas	fluorescens
Pseudomonas	fragi
Pseudomonas	frederiksbergensis
Pseudomonas	gingeri
Pseudomonas	koreensis
Pseudomonas	migulae
Pseudomonas	mohnii
Pseudomonas	monteilii
Pseudomonas	moorei
Pseudomonas	moraviensis
Pseudomonas	plecoglossicida
Pseudomonas	poae
Pseudomonas	putida
Pseudomonas	syringae
Pseudomonas	vancouverensis
Staphylococcus	aureus

8 References

Abia, A.L.K., Alisoltani, A., Keshri, J., Ubomba-Jaswa, E., 2018. Metagenomic analysis of the bacterial communities and their functional profiles in water and sediments of the Apies River, South Africa, as a function of land use. Sci. Total Environ. 616–617, 326–334. https://doi.org/10.1016/j.scitotenv.2017.10.322

ABS, 2022. Water Use on Australian Farms (2020-21) [WWW Document]. Aust. Bur. Stat. URL https://www.abs.gov.au/statistics/industry/agriculture/water-use-australian-farms/latest-release

Ahmed, W., Goonetilleke, A., Gardner, T., 2010. Implications of faecal indicator bacteria for the microbiological assessment of roof-harvested rainwater quality in southeast Queensland, Australia. Can. J. Microbiol. 56, 471–479. https://doi.org/10.1139/W10-037

Akritas, M.G., Murphy, S.A., LaValley, M.P., 1995. The Theil-Sen Estimator With Doubly Censored Data and Applications to Astronomy. J. Am. Stat. Assoc. 90, 170–177.

Allard, S.M., Callahan, M.T., Bui, A., Ferelli, A.M.C., Chopyk, J., Chattopadhyay, S., Mongodin, E.F., Micallef, S.A., Sapkota, A.R., 2019. Creek to Table: Tracking fecal indicator bacteria, bacterial pathogens, and total bacterial communities from irrigation water to kale and radish crops. Sci. Total Environ. 666, 461–471. https://doi.org/10.1016/j.scitotenv.2019.02.179

Allende, A., Monaghan, J., 2015. Irrigation Water Quality for Leafy Crops: A Perspective of Risks and Potential Solutions. Int. J. Environ. Res. Public Health 12, 7457–7477. https://doi.org/10.3390/ijerph120707457

Alum, A., Enriquez, C., Gerba, C.P., 2011. Impact of Drip Irrigation Method, Soil, and Virus Type on Tomato and Cucumber Contamination. Food Environ. Virol. 3, 78–85. https://doi.org/10.1007/s12560-011-9060-y

Antaki, E.M., Vellidis, G., Harris, C., Aminabadi, P., Levy, K., Jay-Russell, M.T., 2016. Low Concentration of *Salmonella enterica* and Generic *Escherichia coli* in Farm Ponds and Irrigation Distribution Systems Used for Mixed Produce Production in Southern Georgia. Foodborne Pathog. Dis. 13, 551–558. https://doi.org/10.1089/fpd.2016.2117

ANZECC & ARMCANZ, 2000. Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Book 1, 314. https://doi.org/ISBN 09578245 0 5

Asaff-Torres, A., Armendáriz-Ruiz, M., Kirchmayr, M., Rodríguez-Heredia, R., Orozco, M., Carlos Mateos-Díaz, J., Figueroa-Yáñez, L., Baqueiro-Peña, I., Verdín, J., Alimentación, E., Desarrollo, A.C., 2017. Rhizospheric Microbiome Profiling of

Capsicum annuum L. Cultivated in Amended Soils by 16S and Internal Transcribed Spacer 2 rRNA Amplicon Metagenome Sequencing. https://doi.org/10.1128/genomeA

Aßhauer, K.P., Wemheuer, B., Daniel, R., Meinicke, P., 2015. Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31, 2882–2884. https://doi.org/10.1093/bioinformatics/btv287

Australia New Zealand Food Standards Code – Standard 4.2.8 – Primary Production and Processing Standard for Leafy Vegetables, 2022.

Australian Horticulture Statistics Handbook 2020/21, 2022.

Balvočiute, M., Huson, D.H., 2017. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics 18, 1–8. https://doi.org/10.1186/s12864-017-3501-4

Barker-Reid, F., Harapas, D., Engleitner, S., Kreidl, S., Holmes, R., Faggian, R., 2009. Persistence of Escherichia coli on injured iceberg lettuce in the field, overhead irrigated with contaminated water. J. Food Prot. 72, 458–64.

Barroso-Bergadà, D., Pauvert, C., Vallance, J., Delière, L., Bohan, D.A., Buée, M., Vacher, C., 2021. Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls. Mol. Ecol. Resour. 21, 762–780. https://doi.org/10.1111/1755-0998.13302

Beach, C., 2018. FDA says poor records stalled outbreak work; feedlot likely source of E. coli | Food Safety News. Food Saf. News.

Belias, A.M., Sbodio, A., Truchado, P., Weller, D., Pinzon, J., Skots, M., Allende, A., Munther, D., Suslow, T., Wiedmann, M., Ivanek, R., 2020. Effect of weather on the die-off of escherichia coli and attenuated salmonella enterica serovar typhimurium on preharvest leafy greens following irrigation with contaminated water. Appl. Environ. Microbiol. 86. https://doi.org/10.1128/AEM.00899-20/SUPPL_FILE/AEM.00899-20-S0001.PDF

Benjamin, L., Atwill, E.R., Jay-Russell, M., Cooley, M., Carychao, D., Gorski, L., Mandrell, R.E., 2013. Occurrence of generic Escherichia coli, E. coli O157 and Salmonella spp. in water and sediment from leafy green produce farms and streams on the Central California coast. Int. J. Food Microbiol. 165, 65–76. https://doi.org/10.1016/j.ijfoodmicro.2013.04.003

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B, 289–300.

Bennett, S.D., Sodha, S. V, Ayers, T.L., Lynch, M.F., Gould, L.H., Tauxe, R. V, 2020. Produce-associated foodborne disease outbreaks , USA , 1998 – 2013.

Berger, C.N., Sodha, S. V., Shaw, R.K., Griffin, P.M., Pink, D., Hand, P., Frankel, G., 2010. Fresh fruit and vegetables as vehicles for the transmission of human pathogens. Environ. Microbiol. 12, 2385–2397. https://doi.org/10.1111/j.1462-2920.2010.02297.x

Bertone, E., Kozak, S., Roiko, A., 2019. Understanding and Modeling the Occurrence of E. coli Blooms in Drinking Water Reservoirs. Water Resour. Res. 55, 10518–10526. https://doi.org/10.1029/2019WR025736

Blaustein, R.A., Pachepsky, Y., Hill, R.L., Shelton, D.R., Whelan, G., 2013. Escherichia coli survival in waters: Temperature dependence. Water Res. 47, 569–578. https://doi.org/10.1016/j.watres.2012.10.027

Bradford, S.A., Harvey, R.W., 2016. Future research needs involving pathogens in groundwater. Hydrogeol. J. 25, 931–938. https://doi.org/10.1007/s10040-016-1501-0

Bradshaw, J.K., Snyder, B.J., Oladeinde, A., Spidle, D., Berrang, M.E., Meinersmann, R.J., Oakley, B., Sidle, R.C., Sullivan, K., Molina, M., 2016. Characterizing relationships among fecal indicator bacteria, microbial source tracking markers, and associated waterborne pathogen occurrence in stream water and sediments in a mixed land use watershed. Water Res. 101, 498–509. https://doi.org/10.1016/j.watres.2016.05.014

Brandl, M.T., Ivanek, R., Zekaj, N., Belias, A., Wiedmann, M., Suslow, T. V., Allende, A., Munther, D.S., 2022. Weather stressors correlate with Escherichia coli and Salmonella enterica persister formation rates in the phyllosphere: a mathematical modeling study. ISME Commun. 2, 1–8. https://doi.org/10.1038/s43705-022-00170-z

Brookes, J.D., Hipsey, M.R., Burch, M.D., Linden, L.G., Ferguson, C.M., Antenucci, J.P., 2005. Relative value of surrogate indicators for detecting pathogens in lakes and reservoirs. Environ. Sci. Technol. 39, 8614–8621. https://doi.org/10.1021/es050821+

Brooks, S.W., Cullor, J.S., Gardner, I.A., 2008. Evaluation of the Colilert® Test for Detection of Verotoxin-Producing Escherichia coli in Water. http://dx.doi.org/10.1300/J096v05n03_05 5, 31–41. https://doi.org/10.1300/J096V05N03_05

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.,

2016a. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B.J., Sankaran, K., Fukuyama, J.A., McMurdie, P.J., Holmes, S.P., 2016b. Bioconductor workflow for microbiome data analysis: From raw reads to community analyses [version 1; referees: 3 approved]. F1000Research 5, 1–49. https://doi.org/10.12688/F1000RESEARCH.8986.1

Callejón, R.M., Rodríguez-Naranjo, M.I., Ubeda, C., Hornedo-Ortega, R., Garcia-Parrilla, M.C., Troncoso, A.M., 2015. Reported foodborne outbreaks due to fresh produce in the United States and European Union: trends and causes. Foodborne Pathog. Dis. 12, 32–8. https://doi.org/10.1089/fpd.2014.1821

Cann, K.F., Thomas, D.R., Salmon, R.L., Wyn-Jones, A.P., Kay, D., 2013. Extreme water-related weather events and waterborne disease. Epidemiol. Infect. 141, 671–686. https://doi.org/10.1017/S0950268812001653

Cao, Y., Fanning, S., Proos, S., Jordan, K., Srikumar, S., 2017. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Front. Microbiol. 8, 1–16. https://doi.org/10.3389/fmicb.2017.01829

Carr, A., Diener, C., Baliga, N.S., Gibbons, S.M., 2019. Use and abuse of correlation analyses in microbial ecology. ISME J. 13, 2647–2655. https://doi.org/10.1038/s41396-019-0459-z

Carroll, S.P., Goonetilleke, A., Hargreaves, M., 2004. Assessment of environmental and public health risk of on-site wastewater treatment systems, in: Proceedings of the 10th National Symposium on Individual and Small Community Sewage Systems - On-Site Wastewater Treatment X. American Society of Agricultural and Biological Engineers, pp. 368–376. https://doi.org/10.13031/2013.15796

Carter, A.M., Pacha, R.E., Clark, G.W., Williams, E.A., 1987. Seasonal occurence of Campylobacter spp. in surface waters and their correlation with standard indicator bacteria. Appl. Environ. Microbiol. 53, 523–526.

CDC, 2022. Summary of Possible Multistate Enteric (Intestinal) Disease Outbreaks in 2017–2020 [WWW Document]. URL https://www.cdc.gov/foodsafety/outbreaks/multistate-outbreaks/annual-summaries/annual-summaries-2017-2020.html#anchor_27531

CDC, 2016. Using Data to Link Foodborne Disease Outbreaks to a Contaminated Source [WWW Document]. URL https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/index.html (accessed 2.6.19).

CDC, 2010. Multistate Outbreak of Human E. coli O145 Infections Linked to Shredded Romaine Lettuce from a Single Processing Facility (FINAL UPDATE) [WWW Document]. URL https://www.cdc.gov/ecoli/2010/shredded-romaine-5-21-10.html (accessed 1.9.19).

CDC, 2006. E. coli Outbreak From Fresh Spinach [WWW Document]. URL https://www.cdc.gov/ecoli/2006/spinach-10-2006.html (accessed 2.6.19).

Cha, Y.K., Park, M.H., Lee, S.H., Kim, J.H., Cho, K.H., 2016. Modeling spatiotemporal bacterial variability with meteorological and watershed land-use characteristics. Water Res. 100, 306–315. https://doi.org/10.1016/j.watres.2016.05.024

Chandran, A., Varghese, S., Kandeler, E., Thomas, A., Hatha, M., Mazumder, A., 2011. An assessment of potential public health risk associated with the extended survival of indicator and pathogenic bacteria in freshwater lake sediments. Int. J. Hyg. Environ. Health 214, 258–264. https://doi.org/10.1016/J.IJHEH.2011.01.002

Chattaway, M.A., Schaefer, U., Tewolde, R., Dallman, T.J., Jenkins, C., 2017. Identification of Escherichia coli and shigella species from whole-genome sequences. J. Clin. Microbiol. 55, 616–623. https://doi.org/10.1128/JCM.01790-16/FORMAT/EPUB

Chen, W., Wilkes, G., Khan, I.U.H., Pintar, K.D.M., Thomas, J.L., Lévesque, C.A., Chapados, J.T., Topp, E., Lapen, D.R., 2018. Aquatic bacterial communities associated with land use and environmental factors in agricultural landscapes using a metabarcoding approach. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.02301

Cheong, S., Lee, C., Song, S.W., Choi, W.C., Lee, C.H., Kim, S.J., 2009. Enteric viruses in raw vegetables and groundwater used for irrigation in South Korea. Appl. Environ. Microbiol. 75, 7745–7751. https://doi.org/10.1128/AEM.01629-09

Cho, K.H., Pachepsky, Y.A., Oliver, D.M., Muirhead, R.W., Park, Y., Quilliam, R.S., Shelton, D.R., 2016. Modeling fate and transport of fecally-derived microorganisms at the watershed scale: State of the science and future opportunities. Water Res. 100, 38–56. https://doi.org/10.1016/j.watres.2016.04.064

Choi, C., Song, I., Stine, S., Pimentel, J., Gerba, C., 2004. Role of irrigation and wastewater reuse: comparison of subsurface irrigation and furrow irrigation. Water Sci. Technol. 50, 61–68. https://doi.org/10.13031/2013.14042

Cito, F., Baldinelli, F., Calistri, P., Giannatale, E. Di, Scavia, G., Orsini, M., Iannetti, S., Sacchini, L., Mangone, I., Candeloro, L., 2014. Outbreak of unusual Salmonella enterica serovar June 2013 to September 2014.

Clarridge, J.E., 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. 17, 840–862. https://doi.org/10.1128/CMR.17.4.840-862.2004

Costa, P.S., Reis, M.P., Ávila, M.P., Leite, L.R., De Araújo, F.M.G., Salim, A.C.M., Oliveira, G., Barbosa, F., Chartone-Souza, E., Nascimento, A.M.A., 2015. Metagenome of a microbial community inhabiting a metal-rich tropical stream sediment. PLoS One 10, 1–21. https://doi.org/10.1371/journal.pone.0119465

Craig, D.L., Fallowfield, H.J., Cromar, N.J., 2004. Use of microcosms to determine persistence of Escherichia coli in recreational coastal water and sediment and validation with in situ measurements. J. Appl. Microbiol. 96, 922–930. https://doi.org/10.1111/J.1365-2672.2004.02243.X

Crawford, W., Baloch, M., 2010. Non-O157 Shiga Toxin-Producing E. coli (Stec) Environmental Assessment Findings And Potential Preventive Control Strategies.

Curriero, F.C., Patz, J.A., Rose, J.B., Lele, S., 2001. The Association Between Extreme Precipitation and Waterborne Disease Outbreaks in the United States , 1948 – 1994. Am. J. Public Health 91, 1194–1199. https://doi.org/10.2105/AJPH.91.8.1194

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K.R., Zeckoski, R.W., Jeong, J., Parajuli, P.B., Saraswat, D., Youssef, M.A., 2015. A Recommended Calibration and Validation Strategy for Hydrologic and Water Quality Models. Trans. ASABE 58, 1705–1719. https://doi.org/10.13031/TRANS.58.10712

Dai, W., Chen, J., Xiong, J., 2019. Concept of microbial gatekeepers: Positive guys? Appl. Microbiol. Biotechnol. 103, 633–641. https://doi.org/10.1007/s00253-018-9522-3

Daly, E., Kolotelo, P., Schang, C., Osborne, C.A., Coleman, R., Deletic, A., McCarthy, D.T., 2013. Escherichia coli concentrations and loads in an urbanised catchment: The Yarra River, Australia. J. Hydrol. 497, 51–61. https://doi.org/10.1016/j.jhydrol.2013.05.024

Danczak, R.E., Johnston, M.D., Kenah, C., Slattery, M., Wilkins, M.J., 2018. Microbial Community Cohesion Mediates Community Turnover in Unperturbed Aquifers. mSystems 3, 1–15. https://doi.org/10.1128/msystems.00066-18

De Deckker, P., Munday, C.I., Brocks, J., O'Loingsigh, T., Allison, G.E., Hope, J., Norman, M., Stuut, J.B.W., Tapper, N.J., van der Kaars, S., 2014. Characterisation of the major dust storm that traversed over eastern Australia in September 2009; a multidisciplinary approach. Aeolian Res. 15, 133–149. https://doi.org/10.1016/j.aeolia.2014.07.003

De Keuckelaere, A., Jacxsens, L., Amoah, P., Medema, G., McClure, P., Jaykus, L.-A.,

Uyttendaele, M., 2015. Zero Risk Does Not Exist: Lessons Learned from Microbial Risk Assessment Related to Use of Water and Safety of Fresh Produce. Compr. Rev. Food Sci. Food Saf. 14, 387–410. https://doi.org/10.1111/1541-4337.12140

Devane, M.L., Moriarty, E.M., Wood, D., Webster-Brown, J., Gilpin, B.J., 2014. The impact of major earthquakes and subsequent sewage discharges on the microbial quality of water and sediments in an urban river. Sci. Total Environ. 485–486, 666–680. https://doi.org/10.1016/j.scitotenv.2014.03.027

Devane, M.L., Weaver, L., Singh, S.K., Gilpin, B.J., 2018. Fecal source tracking methods to elucidate critical sources of pathogens and contaminant microbial transport through New Zealand agricultural watersheds – A review. J. Environ. Manage. 222, 293–303. https://doi.org/10.1016/j.jenvman.2018.05.033

Ditommaso, S., Giacomuzzi, M., Ricciardi, E., Zotti, C.M., 2015. Viability-qPCR for detecting Legionella: Comparison of two assays based on different amplicon lengths. Mol. Cell. Probes 29, 237–243. https://doi.org/10.1016/J.MCP.2015.05.011

Droppo, I.G., Liss, S.N., Williams, D., Nelson, T., Jaskot, C., Trapp, B., 2009. Dynamic existence of waterborne pathogens within river sediment compartments. Implications for water quality regulatory affairs. Environ. Sci. Technol. 43, 1737–1743. https://doi.org/10.1021/es802321w

Duchenne-Moutien, R.A., Neetoo, H., 2021. Climate change and emerging food safety issues: A review. J. Food Prot. https://doi.org/10.4315/JFP-21-141

Erickson, M.C., Webb, C.C., Diaz-Perez, J.C., Phatak, S.C., Silvoy, J.J., Davey, L., Payton, A.S., Liao, J., Ma, L., Doyle, M.P., 2010. Surface and Internalized Escherichia coli O157:H7 on Field-Grown Spinach and Lettuce Treated with Spray-Contaminated Irrigation Water. J. Food Prot. 73, 1023–1029.

F Escapa, I., Huang, Y., Chen, T., Lin, M., Kokaras, A., Dewhirst, F.E., Lemon, K.P., 2020. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8, 65. https://doi.org/10.1186/s40168-020-00841-w

Falardeau, J., Johnson, R.P., Pagotto, F., Wang, S., 2017. Occurrence, characterization, and potential predictors of verotoxigenic Escherichia coli, Listeria monocytogenes, and Salmonella in surface water used for produce irrigation in the Lower Mainland of British Columbia, Canada. PLoS One 12, 1–22. https://doi.org/10.1371/journal.pone.0185437

Farnham, D.J., Lall, U., 2015. Predictive statistical models linking antecedent meteorological conditions and waterway bacterial contamination in urban waterways. Water Res. 76, 143–159. https://doi.org/10.1016/j.watres.2015.02.040

FDA, 2021. Standards for the Growing, Harvesting, Packing, and Holding of Produce for Human Consumption Relating to Agricultural Water.

FDA Food Safety Modernization Act, 2017.

Filippini, G., Bugnot, A.B., Johnston, E.L., Ruszczyk, J., Potts, J., Scanes, P., Ferguson, A., Ostrowski, M., Varkey, D., Dafforn, K.A., 2019. Sediment bacterial communities associated with environmental factors in Intermittently Closed and Open Lakes and Lagoons (ICOLLs). Sci. Total Environ. 693, 133462. https://doi.org/10.1016/j.scitotenv.2019.07.268

Freshcare Ltd, 2020. Food Safety & Quality Standard, Edition 4. ed.

Garner, E., Davis, B.C., Milligan, E., Blair, M.F., Keenum, I., Maile-Moskowitz, A., Pan, J., Gnegy, M., Liguori, K., Gupta, S., Prussin, A.J., Marr, L.C., Heath, L.S., Vikesland, P.J., Zhang, L., Pruden, A., 2021. Next generation sequencing approaches to evaluate water and wastewater quality. Water Res. 194, 116907. https://doi.org/10.1016/j.watres.2021.116907

Gartley, S., Anderson-Coughlin, B., Sharma, M., Kniel, K.E., 2022. Listeria monocytogenes in Irrigation Water: An Assessment of Outbreaks, Sources, Prevalence, and Persistence. Microorganisms 10, 1–13. https://doi.org/10.3390/microorganisms10071319

Gelting, R.J., Baloch, M., 2013. A Systems Analysis of Irrigation Water Quality in Environmental Assessments Related to Foodborne Outbreaks. Aquat. Procedia 1, 130–137. https://doi.org/10.1016/j.aqpro.2013.07.011

Gelting, R.J., Baloch, M.A., Zarate-Bermudez, M., Hajmeer, M.N., Yee, J.C., Brown, T., Yee, B.J., 2015. A systems analysis of irrigation water quality in an environmental assessment of an E. coli O157: H7 outbreak in the United States linked to iceberg lettuce. Agric. Water Manag. 150, 111–118. https://doi.org/10.1016/j.agwat.2014.12.002

Gelting, R.J., Baloch, M.A., Zarate-Bermudez, M.A., Selman, C., 2011. Irrigation water issues potentially related to the 2006 multistate E. coli O157:H7 outbreak associated with spinach. Agric. Water Manag. 98, 1395–1402. https://doi.org/10.1016/j.agwat.2011.04.004

Gerba, C.P., 2000. Assessment of enteric pathogen shedding by bathers during recreational activity and its impact on water quality. Quant. Microbiol. 2, 55–68. https://doi.org/10.1023/A:1010000230103/METRICS

Ghaju Shrestha, R., Tanaka, Y., Malla, B., Bhandari, D., Tandukar, S., Inoue, D., Sei, K.,

Sherchand, J.B., Haramoto, E., 2017. Next-generation sequencing identification of pathogenic bacterial genes and their relationship with fecal indicator bacteria in different water sources in the Kathmandu Valley, Nepal. Sci. Total Environ. 601–602, 278–284. https://doi.org/10.1016/j.scitotenv.2017.05.105

Gibney, K.B., O'Toole, J., Sinclair, M., Leder, K., 2014. Disease burden of selected gastrointestinal pathogens in Australia, 2010. Int. J. Infect. Dis. 28, 176–185. https://doi.org/10.1016/J.IJID.2014.08.006

Gracey, M., Burke, V., Robinson, J., 1982. AEROMONAS-ASSOCIATED GASTROENTERITIS. Lancet 320, 1304–1306. https://doi.org/10.1016/S0140-6736(82)91510-0

Graczyk, T.K., Majewska, A.C., Schwab, K.J., 2008. The role of birds in dissemination of human waterborne enteropathogens. Trends Parasitol. 24, 55–59. https://doi.org/10.1016/j.pt.2007.10.007

Graham, E.B., Crump, A.R., Resch, C.T., Fansler, S., Arntzen, E., Kennedy, D.W., Fredrickson, J.K., Stegen, J.C., 2017. Deterministic influences exceed dispersal effects on hydrologically-connected microbiomes. Environ. Microbiol. 19, 1552–1567. https://doi.org/10.1111/1462-2920.13720

Green, M.L., Karp, P.D., 2006. The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Res. 34, 3687–3697. https://doi.org/10.1093/nar/gkl438

Greene, S.K., Daly, E.R., Talbot, E.A., Demma, L.J., Holzbauer, S., Patel, N.J., Hill, T.A., Walderhaug, M.O., Hoekstra, R.M., Lynch, M.F., Painter, J.A., 2008. Recurrent multistate outbreak of Salmonella Newport associated with tomatoes from contaminated fields, 2005. Epidemiol. Infect. 136, 157–165. https://doi.org/10.1017/S095026880700859X

Guo, F., Ju, F., Cai, L., Zhang, T., 2013. Taxonomic Precision of Different Hypervariable Regions of 16S rRNA Gene and Annotation Methods for Functional Bacterial Groups in Biological Wastewater Treatment. PLoS One 8. https://doi.org/10.1371/journal.pone.0076185

Guo, X.A., van Iersel, M.W., Chen, J.R., Brackett, R.E., Beuchat, L.R., 2002. Evidence of association of salmonellae with tomato plants grown hydroponically in inoculated nutrient solution. Appl. Environ. Microbiol. 68, 3639–3643. https://doi.org/10.1128/aem.68.7.3639-3643.2002

Hach Company, 2018. Hach Method 8203: Alkalinity.

Hajmeer, M., Yee, C., Myers, C., Yee, B., Martinez, D., Kennelly, P., Farrar, J., Cassens, B., Cruze, A., 2007. Environmental Investigation of the Taco John's Escherichia coli O157:H7 Outbreak Linked to Iceberg Lettuce.

Haller, L., Amedegnato, E., Poté, J., Wildi, W., 2009. Influence of Freshwater Sediment Characteristics on Persistence of Fecal Indicator Bacteria. Water. Air. Soil Pollut. 203, 217–227. https://doi.org/10.1007/s11270-009-0005-0

Hansen, S., Messer, T., Mittelstet, A., Berry, E.D., Bartelt-Hunt, S., Abimbola, O., 2020. Escherichia coli concentrations in waters of a reservoir system impacted by cattle and migratory waterfowl. Sci. Total Environ. 705, 135607. https://doi.org/10.1016/j.scitotenv.2019.135607

Harapas, D., Premier, R., Tomkins, B., Franz, P., Ajlouni, S., 2010. Persistence of Escherichia coli on injured vegetable plants. Int. J. Food Microbiol. 138, 232–237. https://doi.org/10.1016/j.ijfoodmicro.2010.01.022

Harmel, R.D., Hathaway, J.M., Wagner, K.L., Wolfe, J.E., Karthikeyan, R., Francesconi, W., McCarthy, D.T., 2016. Uncertainty in monitoring E. coli concentrations in streams and stormwater runoff. J. Hydrol. 534, 524–533. https://doi.org/10.1016/j.jhydrol.2016.01.040

Harris, C.S., Tertuliano, M., Rajeev, S., Vellidis, G., Levy, K., 2018. Impact of storm runoff on Salmonella and Escherichia coli prevalence in irrigation ponds of fresh produce farms in southern Georgia. J. Appl. Microbiol. 124, 910–921. https://doi.org/10.1111/jam.13689

Hauer, T., Komárek, J., 2022. CyanoDB 2.0 - On-line database of cyanobacterial genera.

Herman, K.M., Hall, A.J., Gould, L.H., 2015. Outbreaks attributed to fresh leafy vegetables, United States, 1973-2012. Epidemiol. Infect. 143, 3011–21. https://doi.org/10.1017/S0950268815000047

Hernandez, D.J., David, A.S., Menges, E.S., Searcy, C.A., Afkhami, M.E., 2021. Environmental stress destabilizes microbial networks. ISME J. https://doi.org/10.1038/s41396-020-00882-x

Herren, C.M., McMahon, K.D., 2018. Keystone taxa predict compositional change in microbial communities. Environ. Microbiol. 20, 2207–2217. https://doi.org/10.1111/1462-2920.14257

Herren, C.M., McMahon, K.D., 2017. Cohesion: A method for quantifying the connectivity of microbial communities. ISME J. 11, 2426–2438. https://doi.org/10.1038/ismej.2017.91

Herrig, I.M., Böer, S.I., Brennholt, N., Manz, W., 2015. Development of multiple linear

regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. Water Res. 85, 148–157. https://doi.org/10.1016/j.watres.2015.08.006

Hirneisen, K.A., Sharma, M., Kniel, K.E., 2012. Human Enteric Pathogen Internalization by Root Uptake into Food Crops. Foodborne Pathog. Dis. 9, 396–405. https://doi.org/10.1089/fpd.2011.1044

Hoelzer, K., Moreno Switt, A.I., Wiedmann, M., Boor, K.J., 2018. Emerging needs and opportunities in foodborne disease detection and prevention: From tools to people. Food Microbiol. 75, 65–71. https://doi.org/10.1016/j.fm.2017.07.006

Holvoet, K., Sampers, I., Seynnaeve, M., Uyttendaele, M., 2014. Relationships among hygiene indicators and enteric pathogens in irrigation water, soil and lettuce and the impact of climatic conditions on contamination in the lettuce primary production. Int. J. Food Microbiol. 171, 21–31. https://doi.org/10.1016/j.ijfoodmicro.2013.11.009

Ibekwe, A.M., Ma, J., Murinda, S.E., 2016. Bacterial community composition and structure in an Urban River impacted by different pollutant sources. Sci. Total Environ. 566–567, 1176–1185. https://doi.org/10.1016/j.scitotenv.2016.05.168

Iliev, I., Yahubyan, G., Marhova, M., Apostolova, E., Gozmanova, M., Gecheva, G., Kostadinova, S., Ivanova, A., Baev, V., 2017. Metagenomic profiling of the microbial freshwater communities in two Bulgarian reservoirs. J. Basic Microbiol. 57, 669–679. https://doi.org/10.1002/jobm.201700137

Islam, M.M.M., Sokolova, E., Hofstra, N., 2018. Modelling of river faecal indicator bacteria dynamics as a basis for faecal contamination reduction. J. Hydrol. 563, 1000–1008. https://doi.org/10.1016/j.jhydrol.2018.06.077

Iwai, S., Weinmaier, T., Schmidt, B.L., Albertson, D.G., Poloso, N.J., Dabbagh, K., DeSantis, T.Z., 2016. Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. PLoS One 11, 1–18. https://doi.org/10.1371/journal.pone.0166104

Jablasone, J., Brovko, L.Y., Griffiths, M.W., 2004. A research note: the potential for transfer of Salmonella from irrigation water to tomatoes. J. Sci. Food Agric. 84, 287–289. https://doi.org/10.1002/jsfa.1646

Jagadeesan, B., Gerner-smidt, P., Allard, M.W., Leuillet, S., Winkler, A., Xiao, Y., Cha, S., Vossen, J. Van Der, Tang, S., Katase, M., Mcclure, P., Kimura, B., Ching, L., Chapman, J., Grant, K., 2019. The use of next generation sequencing for improving food safety : Translation into practice 79, 96–115. https://doi.org/10.1016/j.fm.2018.11.005

Jang, J., Hur, H.G., Sadowsky, M.J., Byappanahalli, M.N., Yan, T., Ishii, S., 2017. Environmental Escherichia coli: ecology and public health implications—a review. J. Appl. Microbiol. 123, 570–581. https://doi.org/10.1111/jam.13468

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16, 309–330.

Jenkins, C., Dallman, T.J., Launders, N., Willis, C., Byrne, L., Jorgensen, F., Eppinger, M., Adak, G.K., Aird, H., Elviss, N., Grant, K.A., Morgan, D., McLauchlin, J., 2015. Public health investigation of two outbreaks of shiga toxin-producing Escherichia coli O157 associated with consumption of watercress. Appl. Environ. Microbiol. 81, 3946–3952. https://doi.org/10.1128/AEM.04188-14

Jonaidi Jafari, N., Radfar, M.H., Ghofrani, H., Masoumi Asl, H., 2007. Epidemiological and bacteriological features of the cholera outbreak in Iran (2005). J. Med. Sci. 7, 645–649. https://doi.org/10.3923/jms.2007.645.649

Julian, P., Helsel, D., 2022. NADA2: Data Analysis for Censored Environmental Data.

Jung, J.H., Yoo, C.H., Koo, E.S., Kim, H.M., Na, Y., Jheong, W.H., Jeong, Y.S., 2011. Occurrence of norovirus and other enteric viruses in untreated groundwaters of Korea. J. Water Health 9, 544–555. https://doi.org/10.2166/wh.2011.142

Karanis, P., Kimura, A., 2002. Evaluation of three flocculation methods for the purification of Cryptosporidium parvum oocysts from water samples. Lett. Appl. Microbiol. 34, 444–449. https://doi.org/10.1046/j.1472-765X.2002.01121.x

Karp, D.S., Gennet, S., Kilonzo, C., Partyka, M., Chaumont, N., Atwill, E.R., Kremen, C., 2015. Comanaging fresh produce for nature conservation and food safety. Proc. Natl. Acad. Sci. U. S. A. 112, 11126–11131. https://doi.org/10.1073/pnas.1508435112

Karp, P., Billington, R., Caspi, R., Fulcher, C., Latendresse, M., Kothari, A., Keseler, I., Krummenacker, M., Midford, P., Ong, Q., Ong, W., Paley, S., Subhraveti, P., 2019. The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform.

Kendall, M., 1938. A new measure of rank correlation. Biometrika.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., Van Der Lelie, D., Meyer, F., Stevens, R., Bailey, M.J., Gordon, J.I., Kowalchuk, G.A., Gilbert, J.A., 2012. Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. 2012 306 30, 513–520. https://doi.org/10.1038/nbt.2235

Kruskal, J.B., Wish, M., 1978. Multidimensional scaling 93.

Kumar, R.R., Park, B.J., Cho, J.Y., 2013. Application and environmental risks of livestock manure. J. Korean Soc. Appl. Biol. Chem. 56, 497–503. https://doi.org/10.1007/S13765-013-3184-8/METRICS

Labbé, R.G., Juneja, V.K., 2013. Clostridium perfringens Gastroenteritis. Foodborne Infect. Intox. 99–112. https://doi.org/10.1016/B978-0-12-416041-5.00006-8

Lane, D., 1991. Nucleic acid sequencing techniques in bacterial systematics.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower, C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31, 814–821. https://doi.org/10.1038/nbt.2676

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H., Buckley, T., Cruickshank, R., Dopheide, A., Handley, K., Hermans, S., Kamke, J., Lee, C., MacDiarmid, R., Morales, S., Orlovich, D., Smissen, R., Wood, J., Holdaway, R., 2018. Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. N. Z. J. Ecol. 42. https://doi.org/10.20417/nzjecol.42.9

Leclerc, H., Mossel, D.A.A., Edberg, S.C., Struijk, C.B., 2001. Advances in the Bacteriology of the Coliform Group: Their Suitability as Markers of Microbial Water Safety. Annu. Rev. Microbiol. 55, 201–234. https://doi.org/10.1146/annurev.micro.55.1.201

Li, B., Vellidis, G., Liu, H., Jay-Russell, M., Zhao, S., Hu, Z., Wright, A., Elkins, C.A., 2014. Diversity and antimicrobial resistance of Salmonella enterica isolates from surface water in southeastern United States. Appl. Environ. Microbiol. 80, 6355–6365. https://doi.org/10.1128/AEM.02063-14

Li, J., Wang, Z., Karim, M.R., Zhang, L., 2020. Detection of human intestinal protozoan parasites in vegetables and fruits: a review. Parasites and Vectors 13, 1–19. https://doi.org/10.1186/S13071-020-04255-3/TABLES/6

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest.

Liu, H., Whitehouse, C.A., Li, B., 2018. Presence and Persistence of Salmonella in Water: The Impact on Microbial Quality of Water and Food Safety. Front. Public Heal. 6, 1–13. https://doi.org/10.3389/fpubh.2018.00159

Liu, S., Ren, H., Shen, L., Lou, L., Tian, G., Zheng, P., Hu, B., 2015. pH levels drive bacterial community structure in the Qiantang River as determined by 454 pyrosequencing.

Front. Microbiol. 6, 285. https://doi.org/10.3389/FMICB.2015.00285/ABSTRACT

Llirós, M., Inceoǧlu, Ö., García-Armisen, T., Anzil, A., Leporcq, B., Pigneur, L.M., Viroux, L., Darchambeau, F., Descy, J.P., Servais, P., 2014. Bacterial community composition in three freshwater reservoirs of different alkalinity and trophic status. PLoS One 9, 1–27. https://doi.org/10.1371/journal.pone.0116145

López-Gálvez, F., Truchado, P., Sánchez, G., Aznar, R., Gil, M.I., Allende, A., 2016. Occurrence of enteric viruses in reclaimed and surface irrigation water: relationship with microbiological and physicochemical indicators. J. Appl. Microbiol. 121, 1180–1188. https://doi.org/10.1111/jam.13224

Lothrop, N., Bright, K.R., Sexton, J., Pearce-Walker, J., Reynolds, K.A., Verhougstraete, M.P., 2018. Optimal strategies for monitoring irrigation water quality. Agric. Water Manag. 199, 86–92. https://doi.org/10.1016/j.agwat.2017.12.018

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., Konstantinidis, K.T., 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. U. S. A. 108, 7200–7205. https://doi.org/10.1073/pnas.1015622108

Lyautey, E., Lapen, D.R., Wilkes, G., Mccleary, K., Pagotto, F., Tyler, K., Hartmann, A., Piveteau, P., Rieu, A., Robertson, W.J., Medeiros, D.T., Edge, T.A., Gannon, V., Topp, E., 2007. Distribution and Characteristics of Listeria monocytogenes Isolates from Surface Waters of the South Nation River Watershed, Ontario, Canada. Appl. Environ. Microbiol. 73, 5401–5410. https://doi.org/10.1128/AEM.00354-07

Machado-Moreira, B., Richards, K., Brennan, F., Abram, F., Burgess, C.M., 2019. Microbial Contamination of Fresh Produce: What, Where, and How? Compr. Rev. Food Sci. Food Saf. https://doi.org/10.1111/1541-4337.12487

Mackay, S.J., Ridley, J.P., 1983. Survival and Regrowth of Escherichia coli in Lake Burragorang.

Mälzer, H.J., aus der Beek, T., Müller, S., Gebhardt, J., 2016. Comparison of different model approaches for a hygiene early warning system at the lower Ruhr River, Germany. Int. J. Hyg. Environ. Health 219, 671–680. https://doi.org/10.1016/j.ijheh.2015.06.005

Mateus-Barros, E., Meneghine, A.K., Bagatini, I.L., Fernandes, C.C., Kishi, L.T., Vieira, A.A.H., Sarmento, H., 2019. Comparison of two DNA extraction methods widely used in aquatic microbial ecology. J. Microbiol. Methods 159, 12–17. https://doi.org/10.1016/j.mimet.2019.02.005

McEgan, R., Chandler, J.C., Goodridge, L.D., Danyluk, M.D., 2014. Diversity of Salmonella isolates from central Florida surface waters. Appl. Environ. Microbiol. 80, 6819–6827. https://doi.org/10.1128/AEM.02191-14

McEgan, R., Mootian, G., Goodridge, L.D., Schaffner, D.W., Danyluk, M.D., 2013. Predicting Salmonella populations from biological, chemical, and physical indicators in Florida surface waters. Appl. Environ. Microbiol. 79, 4094–4105. https://doi.org/10.1128/AEM.00777-13

McLaren, M.R., Callahan, B.J., 2021a. Silva 138.1 prokaryotic SSU taxonomic training data formatted for DADA2. https://doi.org/10.5281/ZENODO.4587955

McLaren, M.R., Callahan, B.J., 2021b. Silva 138.1 prokaryotic SSU taxonomic training data formatted for DADA2.

McMurdie, P.J., Holmes, S., 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Mendes Silva, D., Domingues, L., 2015. On the track for an efficient detection of Escherichia coli in water: A review on PCR-based methods. Ecotoxicol. Environ. Saf. 113, 400–411. https://doi.org/10.1016/J.ECOENV.2014.12.015

Mikhail, A.F.W., Jenkins, C., Dallman, T.J., Inns, T., Martín, A.I.C., Fox, A., Cleary, P., Elson, R., Hawker, J., 2018. An outbreak of Shiga toxin-producing Escherichia coli O157:H7 associated with contaminated salad leaves: Epidemiological, genomic and food trace back investigations. Epidemiol. Infect. 146, 187–196. https://doi.org/10.1017/S0950268817002874

Miles, J.M., Sumner, S.S., Boyer, R.R., Williams, R.C., Latimer, J.G., McKinney, J.M., 2009. Internalization of Salmonella enterica Serovar Montevideo into Greenhouse Tomato Plants through Contaminated Irrigation Water or Seed Stock. J. Food Prot. 72, 849–852.

Mody, R.K., Greene, S.A., Gaul, L., Sever, A., Pichette, S., Zambrana, I., Dang, T., Gass, A., Wood, R., Herman, K., Cantwell, L.B., Falkenhorst, G., Wannemuehler, K., Hoekstra, R.M., McCullum, I., Cone, A., Franklin, L., Austin, J., Delea, K., Behravesh, C.B., Sodha, S. V., Yee, J.C., Emanuel, B., Al-Khaldi, S.F., Jefferson, V., Williams, I.T., Griffin, P.M., Swerdlow, D.L., 2011. National outbreak of Salmonella serotype Saintpaul infections: Importance of Texas restaurant investigations in implicating Jalapeño peppers. PLoS One 6, 1–7. https://doi.org/10.1371/journal.pone.0016579

Mohammed, H., Hameed, I.A., Seidu, R., 2017. Random forest tree for predicting fecal indicator organisms in drinking water supply. Proc. 4th Int. Conf. Behav. Econ. Socio-

Cultural Comput. BESC 2017 2018-Janua, 1–6. https://doi.org/10.1109/BESC.2017.8256398

Murphy, M., Jamieson, R., Gordon, R., Stratton, G.W., Madani, A., 2010. Inactivation of Escherichia coli During Storage of Irrigation Water in Agricultural Reservoirs. Can. Water Resour. J. https://doi.org/10.4296/cwrj3501069

Nanayakkara, B.S., 2019. Phenotypic and genotypic characteristics for Escherichia coli strains responsible for bacterial bloom events in Australia.

Nanayakkara, B.S., O'Brien, C.L., Gordon, D.M., 2019. Diversity and distribution of Klebsiella capsules in Escherichia coli. Environ. Microbiol. Rep. 11, 107–117. https://doi.org/10.1111/1758-2229.12710

Narayan, N.R., Weinmaier, T., Laserna-Mendieta, E.J., Claesson, M.J., Shanahan, F., Dabbagh, K., Iwai, S., Desantis, T.Z., 2020. Erratum: Piphillin predicts metagenomic composition and dynamics from DADA2- corrected 16S rDNA sequences (BMC Genomics (2020) 21 (56) DOI: 10.1186/s12864-019-6427-1). BMC Genomics 21, 1–12. https://doi.org/10.1186/s12864-020-6537-9

Nevers, M.B., Byappanahalli, M.N., Nakatsu, C.H., Kinzelman, J.L., Phanikumar, M.S., Shively, D.A., Spoljaric, A.M., 2020. Interaction of bacterial communities and indicators of water quality in shoreline sand, sediment, and water of Lake Michigan. Water Res. 178, 115671. https://doi.org/10.1016/j.watres.2020.115671

Nguyen, H.T.M., Le, Q.T.P., Garnier, J., Janeau, J.L., Rochelle-Newall, E., 2016. Seasonal variability of faecal indicator bacteria numbers and die-off rates in the Red River basin, North Viet Nam. Sci. Reports 2016 61 6, 1–12. https://doi.org/10.1038/srep21644

NHMRC, NRMMC, 2011. Australian Drinking Water Guidelines Paper 6 National Water Quality Management Strategy.

Nho, S.W., Abdelhamed, H., Paul, D., Park, S., Mauel, M.J., Karsi, A., Lawrence, M.L., 2018. Taxonomic and functional metagenomic profile of sediment from a commercial catfish pond in Mississippi. Front. Microbiol. 9, 1–12. https://doi.org/10.3389/fmicb.2018.02855

Nshimyimana, J.P., Freedman, A.J.E., Shanahan, P., Chua, L.C.H., Thompson, J.R., 2017. Variation of Bacterial Communities with Water Quality in an Urban Tropical Catchment. Environ. Sci. Technol. 51, 5591–5601. https://doi.org/10.1021/acs.est.6b04737

NSW Department of Health, 2005. Drinking Water Monitoring Program, October 20. ed.

NSW Landuse 2017 v1.2, 2020.

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara,

R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. De, Durand, S., Beatriz, H., Evangelista, A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Cunha, E.R., Smith, T., Stier, A., Braak, C.J.F. Ter, Weedon, J., 2022. vegan: Community Ecology Package.

Olaimat, A.N., Holley, R.A., 2012. Factors influencing the microbial safety of fresh produce: A review. Food Microbiol. 32, 1–19. https://doi.org/10.1016/j.fm.2012.04.016

Oliveira, M., Viñas, I., Usall, J., Anguera, M., Abadias, M., 2012. Presence and survival of Escherichia coli O157:H7 on lettuce leaves and in soil treated with contaminated compost and irrigation water. Int. J. Food Microbiol. 156, 133–140. https://doi.org/10.1016/j.ijfoodmicro.2012.03.014

Oliver, D.M., Haygarth, P.M., Clegg, C.D., Heathwaite, A.L., 2006. Differential E. coli die-off patterns associated with agricultural matrices. Environ. Sci. Technol. 40, 5710–5716. https://doi.org/10.1021/ES0603249/SUPPL_FILE/ES0603249SI20060217_053451.PDF

Oliver, D.M., Porter, K.D.H., Pachepsky, Y.A., Muirhead, R.W., Reaney, S.M., Coffey, R., Kay, D., Milledge, D.G., Hong, E., Anthony, S.G., Page, T., Bloodworth, J.W., Mellander, P.E., Carbonneau, P.E., McGrane, S.J., Quilliam, R.S., 2016a. Predicting microbial water quality with models: Over-arching questions for managing risk in agricultural catchments. Sci. Total Environ. 544, 39–47. https://doi.org/10.1016/j.scitotenv.2015.11.086

Oliver, D.M., Porter, K.D.H., Pachepsky, Y.A., Muirhead, R.W., Reaney, S.M., Coffey, R., Kay, D., Milledge, D.G., Hong, E., Anthony, S.G., Page, T., Bloodworth, J.W., Mellander, P.E., Carbonneau, P.E., McGrane, S.J., Quilliam, R.S., 2016b. Predicting microbial water quality with models: Over-arching questions for managing risk in agricultural catchments. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2015.11.086

OzFoodNet Working Group, 2022. Monitoring the incidence and causes of disease potentially transmitted by food in Australia: Annual report of the OzFoodNet network, 2017, Communicable Diseases Intelligence.

Pachepsky, Y., Shelton, D., Dorner, S., Whelan, G., 2014. Can *E. coli* or thermotolerant coliform concentrations predict pathogen presence or prevalence in irrigation waters? Crit. Rev. Microbiol. 7828, 1–10. https://doi.org/10.3109/1040841X.2014.954524

Pachepsky, Y., Shelton, D.R., McLain, J.E.T., Patel, J., Mandrell, R.E., 2011. Irrigation waters as a source of pathogenic microorganisms in produce: a review, in: Sparks, D.L.

(Ed.), Advances in Agronomy, Vol 113. Elsevier Academic Press Inc, San Diego, pp. 73–138. https://doi.org/10.1016/b978-0-12-386473-4.00007-5

Pagadala, S., Marine, S.C., Micallef, S.A., Wang, F., Pahl, D.M., Melendez, M. V., Kline, W.L., Oni, R.A., Walsh, C.S., Everts, K.L., Buchanan, R.L., 2015. Assessment of region, farming system, irrigation source and sampling time as food safety risk factors for tomatoes. Int. J. Food Microbiol. 196, 98–108. https://doi.org/10.1016/j.ijfoodmicro.2014.12.005

Palmer, S.R., Gully, P.R., White, J.M., Pearson, A.D., Suckling, W.G., Jones, D.M., Rawes, J.C.L., Penner, J.L., 1983. WATER-BORNE OUTBREAK OF CAMPYLOBACTER GASTROENTERITIS. Lancet 321, 287–290. https://doi.org/10.1016/S0140-6736(83)91698-7

Park, S., Navratil, S., Gregory, A., Bauer, A., Srinath, I., Jun, M., Szonyi, B., Nightingale, K., Anciso, J., Ivanek, R., 2013. Generic Escherichia coli contamination of spinach at the preharvest stage: Effects of farm management and environmental factors. Appl. Environ. Microbiol. 79, 4347–4358. https://doi.org/10.1128/AEM.00474-13

Park, Y., Pachepsky, Y., Hong, E.M., Shelton, D., Coppock, C., 2017. Escherichia coli release from streambed to water column during baseflow periods: A modeling study. J. Environ. Qual. 46, 219–226. https://doi.org/10.2134/jeq2016.03.0114

Paton, J.C., Paton, A.W., 1998. Pathogenesis and diagnosis of Shiga toxin-producing Escherichia coli infections. Clin. Microbiol. Rev. 11, 450–479.

Paule-Mercado, M.A., Ventura, J.S., Memon, S.A., Jahng, D., Kang, J.H., Lee, C.H., 2016. Monitoring and predicting the fecal indicator bacteria concentrations from agricultural, mixed land use and urban stormwater runoff. Sci. Total Environ. 550, 1171–1181. https://doi.org/10.1016/j.scitotenv.2016.01.026

Petit, F., Clermont, O., Delannoy, S., Servais, P., Gourmelon, M., Fach, P., Oberlé, K., Fournier, M., Denamur, E., Berthe, T., 2017. Change in the structure of Escherichia coli population and the pattern of virulence genes along a rural aquatic continuum. Front. Microbiol. 8, 1–14. https://doi.org/10.3389/fmicb.2017.00609

Peto, R., Peto, J., 1972. Asymptotically Efficient Rank Invariant Test Procedures. J. R. Stat. Soc. Series A, 185. https://doi.org/10.2307/2344317

Phelan, B., 2019. Pykes Creek reopens [WWW Document]. South. Rural Water. URL http://www.srw.com.au/14532-2/

Poudel, R., Jumpponen, A., Schlatter, D.C., Paulitz, T.C., McSpadden Gardener, B.B.,

Kinkel, L.L., Garrett, K.A., 2016. Microbiome networks: A systems framework for identifying candidate microbial assemblages for disease management. Phytopathology 106, 1083–1096. https://doi.org/10.1094/PHYTO-02-16-0058-FI

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., Slade, M.B., 2005. Phenotypic and genotypic characterization of encapsulated Escherichia coli isolated from blooms in two Australian lakes. Environ. Microbiol. 7, 631–640. https://doi.org/10.1111/j.1462-2920.2005.00729.x

Psarras, G., Chartzoulakis, K., Kasapakis, I., Kloppmann, W., 2014. Effect of Different Irrigation Techniques and Water Qualities on Yield, Fruit Quality and Health Risks of Tomato Plants, in: Braun, P., Stoll, M., Zinkernagel, J. (Eds.), Vii International Symposium on Irrigation of Horticultural Crops. Int Soc Horticultural Science, Leuven 1, pp. 601–608.

Qiu, Y.-Q., 2013. KEGG Pathway Database BT - Encyclopedia of Systems Biology, in: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), . Springer New York, New York, NY, NY, pp. 1068–1069. https://doi.org/10.1007/978-1-4419-9863-7_472

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41. https://doi.org/10.1093/nar/gks1219

Quintanilla Portillo, J., Cheng, X., Belias, A.M., Weller, D.L., Wiedmann, M., Stasiewicz, M.J., 2022. A Validated Preharvest Sampling Simulation Shows that Sampling Plans with a Larger Number of Randomly Located Samples Perform Better than Typical Sampling Plans in Detecting Representative Point-Source and Widespread Hazards in Leafy Green Fields. Appl. Environ. Microbiol. 88, e0101522. https://doi.org/10.1128/AEM.01015-22

R Core Team, 2021. R: A Language and Environment for Statistical Computing.

Rahn, A., Drummelsmith, J., Whitfield, C., 1999. Conserved organization in the cps gene clusters for expression of Escherichia coli group 1 K antigens: Relationship to the colanic acid biosynthesis locus and the cps genes from Klebsiella pneumoniae. J. Bacteriol. 181, 2307–2313.

Rajwar, A., Srivastava, P., Sahgal, M., 2016. Microbiology of Fresh Produce: Route of Contamination, Detection Methods, and Remedy. Crit. Rev. Food Sci. Nutr. 56, 2383–2390. https://doi.org/10.1080/10408398.2013.841119

Ravaliya, K., Gentry-Shields, J., Garcia, S., Heredia, N., de Aceituno, A.F., Bartz, F.E., Leon,

J.S., Jaykus, L.A., 2014. Use of bacteroidales microbial source tracking to monitor fecal contamination in fresh produce production. Appl. Environ. Microbiol. 80, 612–617. https://doi.org/10.1128/AEM.02891-13/FORMAT/EPUB

Reitz, A., Hemric, E., Hall, K.K., 2021. Evaluation of a multivariate analysis modeling approach identifying sources and patterns of nonpoint fecal pollution in a mixed use watershed. J. Environ. Manage. 277, 111413. https://doi.org/10.1016/J.JENVMAN.2020.111413

Rendueles, O., Garcia-garcer, M., Neron, B., Touchon, M., Rocha, E., 2017. Abundance and co-occurence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLoS Pathog 13, 1–30.

Riedel, T.E., Thulsiraj, V., Zimmer-Faust, A.G., Dagit, R., Krug, J., Hanley, K.T., Adamek, K., Ebentier, D.L., Torres, R., Cobian, U., Peterson, S., Jay, J.A., 2015. Long-term monitoring of molecular markers can distinguish different seasonal patterns of fecal indicating bacteria sources. Water Res. 71, 227–243. https://doi.org/10.1016/j.watres.2014.12.037

Rizzatti, G., Lopetuso, L.R., Gibiino, G., Binda, C., Gasbarrini, A., 2017. Proteobacteria: A common factor in human diseases. Biomed Res. Int. 2017. https://doi.org/10.1155/2017/9351507

Roberto, A.A., Van Gray, J.B., Leff, L.G., 2018. Sediment bacteria in an urban stream: Spatiotemporal patterns in community composition. Water Res. 134, 353–369. https://doi.org/10.1016/j.watres.2018.01.045

Rock, C., Rivera, B., Gerba, C.P., 2015. Microbial Source Tracking, Environmental Microbiology: Third Edition. Elsevier Inc. https://doi.org/10.1016/B978-0-12-394626-3.00014-4

Rock, C.M., Brassill, N., Dery, J.L., Carr, D., McLain, J.E., Bright, K.R., Gerba, C.P., 2019. Review of water quality criteria for water reuse and risk-based implications for irrigated produce under the FDA Food Safety Modernization Act, produce safety rule. Environ. Res. 172, 616–629. https://doi.org/10.1016/j.envres.2018.12.050

Rock, C.M., Gerba, C.P., Bright, K.R., Tamimi, A., 2016. Final Project Report: Evaluation of risk-based water quality sampling strategies for the fresh produce industry.

Rodríguez-López, P., Rodríguez-Herrera, J.J., Vázquez-Sánchez, D., Cabo, M.L., 2018. Current Knowledge on Listeria monocytogenes Biofilms in Food-Related Environments: Incidence, Resistance to Biocides, Ecology and Biocontrol. Foods 2018, Vol. 7, Page 85 7, 85. https://doi.org/10.3390/FOODS7060085

Rohwer, R.R., Hamilton, J.J., Newton, R.J., McMahon, K.D., 2018. TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. mSphere 3, 1–14. https://doi.org/10.1128/msphere.00327-18

Röttjers, L., Faust, K., 2018. From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiol. Rev. https://doi.org/10.1093/femsre/fuy030

Rusiñol, M., Martínez-Puchol, S., Timoneda, N., Fernández-Cassi, X., Pérez-Cataluña, A., Fernández-Bravo, A., Moreno-Mesonero, L., Moreno, Y., Alonso, J.L., Figueras, M.J., Abril, J.F., Bofill-Mas, S., Girones, R., 2020. Metagenomic analysis of viruses, bacteria and protozoa in irrigation water. Int. J. Hyg. Environ. Health 224. https://doi.org/10.1016/j.ijheh.2019.113440

Sassi, H.P., van Ogtrop, F., Morrison, C.M., Zhou, K., Duan, J.G., Gerba, C.P., 2020. Sediment re-suspension as a potential mechanism for viral and bacterial contaminants. J. Environ. Sci. Heal. - Part A Toxic/Hazardous Subst. Environ. Eng. 55, 1398–1405. https://doi.org/10.1080/10934529.2020.1796118

Schmid-Hempel, P., Frank, S.A., 2007. Pathogenesis, Virulence, and Infective Dose. PLoS Pathog. 3, 1372–1373. https://doi.org/10.1371/JOURNAL.PPAT.0030147

Sequino, G., Valentino, V., Torrieri, E., De Filippis, F., 2022. Specific Microbial Communities Are Selected in Minimally-Processed Fruit and Vegetables according to the Type of Product. Foods 11. https://doi.org/10.3390/foods11142164

Servais, P., Prats, J., Passerat, J., Garcia-Armisen, T., 2009. Abundance of culturable versus viable Escherichia coli in freshwater. Can. J. Microbiol. 55, 905–909. https://doi.org/10.1139/W09-043

Shang, Y., Wu, X., Wang, X., Wei, Q., Ma, S., Sun, G., Zhang, Huanxin, Wang, L., Dou, H., Zhang, Honghai, 2022. Factors affecting seasonal variation of microbial community structure in Hulun Lake, China. Sci. Total Environ. 805, 150294. https://doi.org/10.1016/j.scitotenv.2021.150294

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/J.1538-7305.1948.TB01338.X

Shilei, Z., Yue, S., Tinglin, H., Ya, C., Xiao, Y., Zizhen, Z., Yang, L., Zaixing, L., Jiansheng, C., Xiao, L., 2020. Reservoir water stratification and mixing affects microbial community structure and functional community composition in a stratified drinking reservoir. J. Environ. Manage. 267, 110456. https://doi.org/10.1016/j.jenvman.2020.110456

Sibanda, T., Chigor, V.N., Okoh, A.I., 2013. Seasonal and spatio-temporal distribution of

faecal-indicator bacteria in Tyume River in the Eastern Cape Province, South Africa. Environ. Monit. Assess. 185, 6579–6590. https://doi.org/10.1007/S10661-012-3048-4/FIGURES/8

Sinclair, M., 2019. Discussion Paper - Identification and Management of Environmental E. coli blooms. WaterRA Project #1101.

Smolders, A., Rolls, R.J., Ryder, D., Watkinson, A., Mackenzie, M., 2015. Cattle-derived microbial input to source water catchments: An experimental assessment of stream crossing modification. J. Environ. Manage. 156, 143–149. https://doi.org/10.1016/J.JENVMAN.2015.03.052

Söderström, A., Österberg, P., Lindqvist, A., Jönsson, B., Lindberg, A., Blide Ulander, S., Welinder-Olsson, C., Löfdahl, S., Kaijser, B., De Jong, B., Kühlmann-Berenzon, S., Boqvist, S., Eriksson, E., Szanto, E., Andersson, S., Allestam, G., Hedenström, I., Ledet Muller, L., Andersson, Y., 2008. A Large Escherichia coli O157 Outbreak in Sweden Associated with Locally Produced Lettuce. Foodborne Pathog. Dis. 5, 339–349. https://doi.org/10.1089/fpd.2007.0065

Sojobi, A.O., Zayed, T., 2022. Impact of sewer overflow on public health: A comprehensive scientometric analysis and systematic review. Environ. Res. 203, 111609. https://doi.org/10.1016/J.ENVRES.2021.111609

Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R., Palmer, C.J., 2000. Sources of Escherichia coli in a coastal subtropical environment. Appl. Environ. Microbiol. 66, 230–237. https://doi.org/10.1128/AEM.66.1.230-237.2000

Solomon, E.B., Potenski, C.J., Matthews, K.R., 2002. Effect of irrigation method on transmission to and persistence of Escherichia coli O157 : H7 on lettuce. J. Food Prot. 65, 673–676.

St-Pierre, K., Lévesque, S., Frost, E., Carrier, N., Arbeit, R.D., Michaud, S., 2009. Thermotolerant coliforms are not a good surrogate for Campylobacter spp. in environmental water. Appl. Environ. Microbiol. 75, 6736–44. https://doi.org/10.1128/AEM.00486-09

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B., Sadowsky, M.J., 2016. Sediments and Soils Act as Reservoirs for Taxonomic and Functional Bacterial Diversity in the Upper Mississippi River. Microb. Ecol. 71, 814–824. https://doi.org/10.1007/s00248-016-0729-5

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B., Sadowsky, M.J., 2014. Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River.

Front. Microbiol. 5, 1–13. https://doi.org/10.3389/fmicb.2014.00524

Staley, C., Sadowsky, M.J., 2016. Application of metagenomics to assess microbial communities in water and other environmental matrices. J. Mar. Biol. Assoc. United Kingdom 96, 121–129. https://doi.org/10.1017/S0025315415001496

Stea, E.C., Purdue, L.M., Jamieson, R.C., Yost, C.K., Hansen, L.T., 2015. Comparison of the prevalences and diversities of Listeria species and Listeria monocytogenes in an urban and a rural agricultural watershed. Appl. Environ. Microbiol. 81, 3812–3822. https://doi.org/10.1128/AEM.00416-15/SUPPL_FILE/ZAM999116293SO1.PDF

Steele, M., Odumeru, J., 2004. Irrigation water as source of foodborne pathogens on fruit and vegetables. J. Food Prot. 67, 2839–2849.

Streeter, K., Katouli, M., 2016. Pseudomonas aeruginosa: A review of their Pathogenesis and Prevalence in Clinical Settings and the Environment. Infect Epidemiol Med. 2016 Winter 2, 25–32. https://doi.org/10.18869/modares.iem.2.1.25

Sulaiman, K., Hakim Ismail, L., Adib Mohammad Razi, M., Shalahuddin Adnan, M., Ghazali, R., 2019. Water Quality Classification Using an Artificial Neural Network (ANN). IOP Conf. Ser. Mater. Sci. Eng. 601. https://doi.org/10.1088/1757-899X/601/1/012005

Sun, Y., Li, X., Liu, J., Yao, Q., Jin, J., Liu, X., Wang, G., 2019. Comparative analysis of bacterial community compositions between sediment and water in different types of wetlands of northeast China. J. Soils Sediments 19, 3083–3097. https://doi.org/10.1007/S11368-019-02301-X/FIGURES/5

Telias, A., White, J.R., Pahl, D.M., Ottesen, A.R., Walsh, C.S., 2011. Bacterial community diversity and variation in spray water sources and the tomato fruit surface. https://doi.org/10.1186/1471-2180-11-81

Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal Escherichia coli. Nat. Rev. Microbiol. 8, 207–217. https://doi.org/10.1038/nrmicro2298

Therneau, T.M., 2022. A Package for Survival Analysis in R.

Thomas, V., 2012. Sensitivity and Resistance of Protozoa to Microbicides. Russell, Hugo Ayliffe's Princ. Pract. Disinfect. Preserv. Steriliz. 155–177. https://doi.org/10.1002/9781118425831.CH8

Touchon, M., Perrin, A., de Sousa, J.A.M., Vangchhia, B., Burn, S., O'Brien, C.L., Denamur, E., Gordon, D., Rocha, E.P., 2020. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli, PLOS Genetics.

https://doi.org/10.1371/journal.pgen.1008866

Tousi, E.G., Duan, J.G., Gundy, P.M., Bright, K.R., Gerba, C.P., 2021. Evaluation of E. coli in sediment for assessing irrigation water quality using machine learning. Sci. Total Environ. 799, 149286. https://doi.org/10.1016/j.scitotenv.2021.149286

Truchado, P., Hernandez, N., Gil, M.I., Ivanek, R., Allende, A., 2017. Correlation between E. coli levels and the presence of foodborne pathogens in surface irrigation water: Establishment of a sampling program. Water Res. 128, 226–233. https://doi.org/10.1016/j.watres.2017.10.041

Truchado, P., Lopez-Galvez, F., Gil, M.I., Pedrero-Salcedo, F., Alarcón, J.J., Allende, A., 2016. Suitability of different Escherichia coli enumeration techniques to assess the microbial quality of different irrigation water sources. Food Microbiol. 58, 29–35. https://doi.org/10.1016/j.fm.2016.03.006

University of Arizona, 2015. AgWater App.

US EPA, 2005. Microbial Source Tracking Guide Document.

Uyttendaele, M., Jaykus, L.A., Amoah, P., Chiodini, A., Cunliffe, D., Jacxsens, L., Holvoet, K., Korsten, L., Lau, M., McClure, P., Medema, G., Sampers, I., Jasti, P.R., Rao Jasti, P., 2015. Microbial Hazards in Irrigation Water: Standards, Norms, and Testing to Manage Use of Water in Fresh Produce Primary Production. Compr. Rev. Food Sci. Food Saf. 14, 336–356. https://doi.org/10.1111/1541-4337.12133

Větrovský, T., Baldrian, P., 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS One 8, 1–10. https://doi.org/10.1371/journal.pone.0057923

Vincent, K., Starrs, D., Wansink, V., Waters, N., Lal, A., 2022. Relationships between extreme flows and microbial contamination in inland recreational swimming areas. J. Water Health 20, 781–793. https://doi.org/10.2166/WH.2022.294

Wang, L., Zhang, J., Li, H., Yang, H., Peng, C., Peng, Z., Lu, L., 2018. Shift in the microbial community composition of surface water and sediment along an urban river. Sci. Total Environ. 627, 600–612. https://doi.org/10.1016/j.scitotenv.2018.01.203

Wang, Y.G., Kuhnert, P., Henderson, B., 2011. Load estimation with uncertainties from opportunistic sampling data - A semiparametric approach. J. Hydrol. 396, 148–157. https://doi.org/10.1016/j.jhydrol.2010.11.003

Ward, M., Dhingra, R., Remais, J. V., Chang, H.H., Johnston, L.M., Jaykus, L.A., Leon, J., 2015. Associations between weather and microbial load on fresh produce prior to

harvest. J. Food Prot. 78, 849–854. https://doi.org/10.4315/0362-028X.JFP-14-381

Warish, A., Triplett, C., Gomi, R., Gyawali, P., Hodgers, L., Toze, S., 2015. Assessment of Genetic Markers for Tracking the Sources of Human Wastewater Associated Escherichia coli in Environmental Waters. Environ. Sci. Technol. 49, 9341–9346. https://doi.org/10.1021/acs.est.5b02163

Weller, D., Belias, A., Green, H., Roof, S., Wiedmann, M., 2020a. Landscape, Water Quality, and Weather Factors Associated With an Increased Likelihood of Foodborne Pathogen Contamination of New York Streams Used to Source Water for Produce Production. Front. Sustain. Food Syst. 3. https://doi.org/10.3389/fsufs.2019.00124

Weller, D., Brassill, N., Rock, C., Ivanek, R., Mudrak, E., Roof, S., Ganda, E., Wiedmann, M., 2020b. Complex Interactions Between Weather, and Microbial and Physicochemical Water Quality Impact the Likelihood of Detecting Foodborne Pathogens in Agricultural Water. Front. Microbiol. 11, 1–20. https://doi.org/10.3389/fmicb.2020.00134

Wen, X., Chen, F., Lin, Y., Zhu, H., Yuan, F., Kuang, D., Jia, Z., Yuan, Z., 2020. Microbial Indicators and Their Use for Monitoring Drinking Water Quality—A Review. Sustain. 2020, Vol. 12, Page 2249 12, 2249. https://doi.org/10.3390/SU12062249

WHO, 2019. Water, sanitation, hygiene and health: a primer for health professionals. Geneva.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Wickham, H., Francois, R., Henry, L., Muller, K., 2022. dplyr: A Grammar of Data Manipulation.

Wilcock, R.J., Nash, D., Schmidt, J., Larned, S.T., Rivers, M.R., Feehan, P., 2011. Inputs of nutrients and fecal bacteria to freshwaters from irrigated agriculture: Case studies in Australia and New Zealand. Environ. Manage. 48, 198–211. https://doi.org/10.1007/s00267-011-9644-1

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. Biometrics Bull. 1, 80. https://doi.org/10.2307/3001968

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., Neumann, N., Ruecker, N., Topp, E., Lapen, D.R., 2009. Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. Water Res. 43, 2209–2223. https://doi.org/10.1016/j.watres.2009.01.033

Wither, A., Rehfisch, M., Austin, G., 2005. The impact of bird populations on the

microbiological quality of bathing waters. Water Sci. Technol. 51, 199–207. https://doi.org/10.2166/WST.2005.0592

Withers, P.J.A., Jordan, P., May, L., Jarvie, H.P., Deal, N.E., 2014. Do septic tank systems pose a hidden threat to water quality? Front. Ecol. Environ. 12, 123–130. https://doi.org/10.1890/130131

Won, G., Schlegel, P.J., Schrock, J.M., LeJeune, J.T., 2013. Absence of Direct Association between Coliforms and Escherichia coli in Irrigation Water and on Produce. J. Food Prot. 76, 959–66. https://doi.org/10.4315/0362-028X.JFP-12-512

Wood, S., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. 73, 3–36.

Wu, J., Long, S.C., Das, D., Dorner, S.M., 2011. Are microbial indicators and pathogens correlated? A statistical analysis of 40 years of research. J. Water Health 9, 265–278. https://doi.org/10.2166/wh.2011.117

Xie, Y., Liu, X., Wei, H., Chen, X., Gong, N., Ahmad, S., Lee, T., Ismail, S., Ni, S.Q., 2022. Insight into impact of sewage discharge on microbial dynamics and pathogenicity in river ecosystem. Sci. Reports 2022 121 12, 1–12. https://doi.org/10.1038/s41598-022-09579-x

Zhang, J., Qiu, H., Li, X., Niu, J., Nevers, M.B., Hu, X., Phanikumar, M.S., 2018. Real-Time Nowcasting of Microbiological Water Quality at Recreational Beaches: A Wavelet and Artificial Neural Network-Based Hybrid Modeling Approach. Environ. Sci. Technol. 52, 8446–8455. https://doi.org/10.1021/ACS.EST.8B01022/ASSET/IMAGES/LARGE/ES-2018-01022W_0003.JPEG

Zhang, L., Delgado-Baquerizo, M., Shi, Y., Liu, X., Yang, Y., Chu, H., 2021. Co-existing water and sediment bacteria are driven by contrasting environmental factors across glacier-fed aquatic systems. Water Res. 198, 117139. https://doi.org/10.1016/J.WATRES.2021.117139

Zhou, K., Sassi, H.P., Morrison, C.M., Duan, J.G., Gerba, C.P., 2017. Resuspension of Escherichia coli and MS2 Bacteriophage from Bed Sediment in Irrigation Canals. J. Irrig. Drain. Eng. 143, 9. https://doi.org/10.1061/(asce)ir.1943-4774.0001169

Zimmer-Faust, A.G., Thulsiraj, V., Marambio-Jones, C., Cao, Y., Griffith, J.F., Holden, P.A., Jay, J.A., 2017. Effect of freshwater sediment characteristics on the persistence of fecal indicator bacteria and genetic markers within a Southern California watershed. Water Res. 119, 1–11. https://doi.org/10.1016/j.watres.2017.04.028