# PREDICTING WATER QUALITY USING MACHINE LEARNING

***Bridget Armstrong (ESR), Theo Sarris (ESR), Louise Weaver (ESR), Helen Morris (University of Canterbury), Judith Webber (ESR), David Wood (ESR)***

---

**ABSTRACT**

Groundwater is utilised for vital services, including drinking water, irrigation, and source water. It is imperative that groundwater resources are protected from increasing anthropogenic activities and that close monitoring of groundwater aquifers is undertaken. Currently, the status of groundwater is monitored by testing specific water quality parameters, including major ions, nitrate-nitrogen, ammonia-nitrogen, silica, iron, and manganese. Faecal indicator bacteria are also monitored as a proxy for pathogenic bacteria from faecal contamination. Currently, these tests are all reactive and indicate a past problem if 'issues' are detected. There is a need for a fast, proactive method to assess the status of groundwater.

While groundwater systems can contain pathogenic bacteria, not all bacteria in this environment are harmful. Nonpathogenic bacteria communities in groundwater aquifers play an essential role in nutrient processing. The composition of these bacterial communities adapts to even subtle changes in the chemical composition of a groundwater system. Changes in bacterial communities could potentially be used as an early warning of a change in water quality, such as nitrate contamination.

Chemical parameters and environmental DNA (eDNA) were compiled from groundwater wells across New Zealand. The machine learning modellings Classification and Regression Training (CART) and Random Forest were used to train various regression algorithms. Several bacterial Phyla were indicated as significant predictors of nitrate levels, i.e., Proteobacteria and Thaumarchaeota.

**KEYWORDS**

**Groundwater, contamination, nitrate, eDNA, machine learning, Random Forest**

**PRESENTER PROFILE**

Judith is a groundwater microbiology scientist in the Groundwater Group at ESR. She is currently extending her skills into bioinformatics learning pipelines to process amplicon sequencing data and perform machine learning techniques for environmental samples.

## INTRODUCTION

Groundwater provides vital services such as drinking water, irrigation, and source water. These critical services are used to cultivate high-yield crops, sustain human

and animal health, and provide energy, material production, and functional infrastructure (Akhtar et al., 2021). As the world's population increases, there is not only increased demand for freshwater resources but the potential for contamination of these freshwater resources. Groundwater is especially vulnerable to contamination in areas where population density is high and human use of the land is intensive. Chemical usage and disposal of waste have the potential to move from the surface waters into the groundwater aquifers. Therefore, resources such as groundwater are monitored for harmful contaminants and pathogens to protect human and animal health.

Currently, groundwater quality is monitored using chemical testing such as nitrate, phosphate, chloride, iron etc., and for faecal contamination, i.e., *Escherichia coli* (an indicator of pathogens). Current testing methods are not performed in real-time, indicating a past problem if issues are detected, and the contamination is potentially already in the community. An example is the Havelock North Campylobacter outbreak (New Zealand) in 2016, where over 8000 people became ill after contaminated drinking water (Gilpin et al., 2020).

There is a need for a fast, proactive method to assess groundwater status to identify groundwater contamination before it is used within the community. Groundwater aquifers are home to a wide variety of microbial communities that utilise contaminants as energy (carbon) sources, and these communities help to remove harmful pollutants from groundwater sources. Microbes respond quickly to subtle chemical changes in their environment, and changes in their diversity could be used as an early warning of upcoming changes in water quality.

Microbial diversity can be screened using environmental DNA (eDNA) (Salis et al., 2017). The monitoring of eDNA in groundwater over time could be used to track and assess the microbes present in groundwater systems. Because microbes are susceptible to chemical changes within their environment, changes in microbial diversity could indicate potential contamination of the groundwater aquifer.

This paper compiles chemical and eDNA parameters from groundwater aquifers across New Zealand. The aim was to model microbial diversity with water chemistry using the predictive modelling tools CART and Random Forest to find key organisms that could predict contaminant presence, i.e., nitrate. The detection of specific microbes or changes in their abundance could ultimately be used for in-line sensors, automated to signal water quality changes.

# MATERIALS AND METHODS

## SITES

Forty-nine existing wells from across the North and South Islands of New Zealand were used in this study (Figure 1). Nine wells were selected from the Hawkes Bay region, five from the Nelson region, thirty from the Canterbury region, and five from the Southland region.

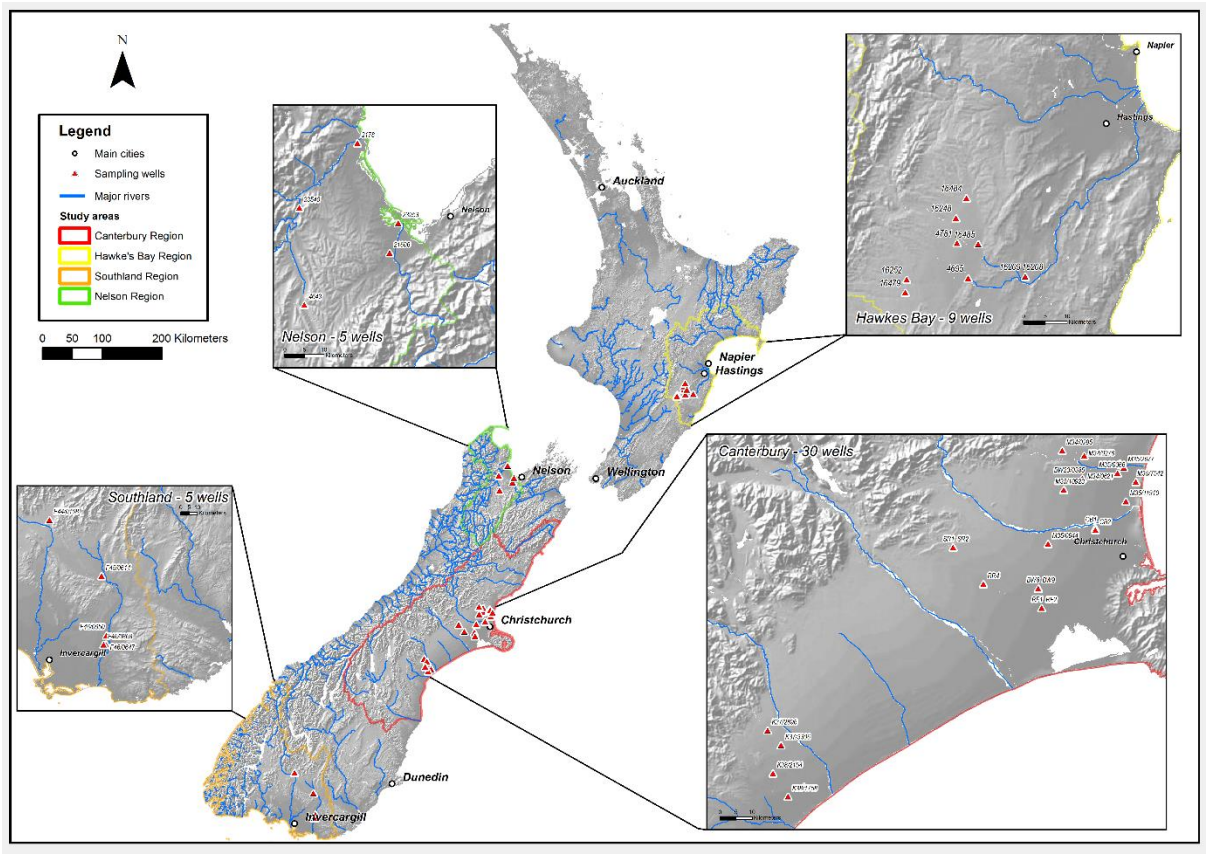


*Figure 1: Location of wells for field sampling*

## GROUNDWATER SAMPLING

Groundwater samples were collected from each well using a Twister pump (plastic impellor) positioned approximately 1 m from the bottom of the well. Wells were purged for at least three well volumes.

Samples were collected and sent to Hill Laboratories in Christchurch or Hamilton for chemical analyses. Analyses included dissolved organic carbon (DOC), dissolved reactive phosphorus (DRP), nitrite-nitrogen (NO2-N), nitrate-nitrogen (NO3-N), ammoniacal-nitrogen (NH4–N), total dissolved nitrogen (TDN), and total dissolved phosphorus (TDP).

Field monitoring was performed onsite and included measuring dissolved oxygen (DO) using a calibrated dissolved oxygen meter (TPS WP-82, TPS Pty Ltd, Australia), and in situ measurements (conductivity, water temperature and pH) taken using a calibrated water quality instrument (TPS WP-81, TPS Pty Ltd, Australia).

Groundwater samples (1 l) were transported to the lab on ice and stored at 4°C overnight. Filters were frozen at -80°C until eDNA extraction. DNA was extracted using the Qiagen DNeasy Power Water® extraction kit (Bio-Strategy Ltd., Auckland, New Zealand). The DNA concentration was measured using a NanoDrop® ND-1000 UV-Vis Spectrophotometer (ThermoFisher Scientific).

## EDNA

Extracted DNA was sent to New Zealand Genomes Ltd (Auckland, New Zealand) for MiSeq metagenomic sequencing (2 x 250 bp PE) using the Illumina MiSeq platform (Illumina, USA) (Caporaso et al., 2012). The prokaryotic 16s rRNA gene was targeted, the returned sequences were quality checked, trimmed to remove barcodes and primers, and the 16S taxonomy was assigned using the DADA2 (v1.10.1) pipeline (Callahan et al., 2016).

## MODELLING

A subset of the chemical and 16S eDNA dataset was modelled to assess the capability of machine learning to predict nitrate levels using bacterial phyla. The two most important bacterial phyla were chosen and modelled against the WHO MAV (Maximum Allowable Value) nitrate classification system (Table 1). The models included CART and Random Forest (v4.6-14). Random Forest was trained with 1000 trees using the training dataset.

*Table 1 WHO MAV nitrate classification system*

| NZ Drinking Water Standard | Nitrate-N (mg/l) | Category |
|---|---|---|
| WHO MAV | 0 to 1 | Pristine |
| | 1 to 11.3 | Below MAV |
| | >11.3 | Above MAV |

# RESULTS

## ENVIRONMENTAL DNA

Seventeen bacterial phyla were present at a level over 1 % relative abundance, with the most abundant Phyla being Proteobacteria (Figure 2). Across all wells, common Phyla were present, so a significant difference in bacterial diversity was not seen, but the relative abundances of bacteria varied.

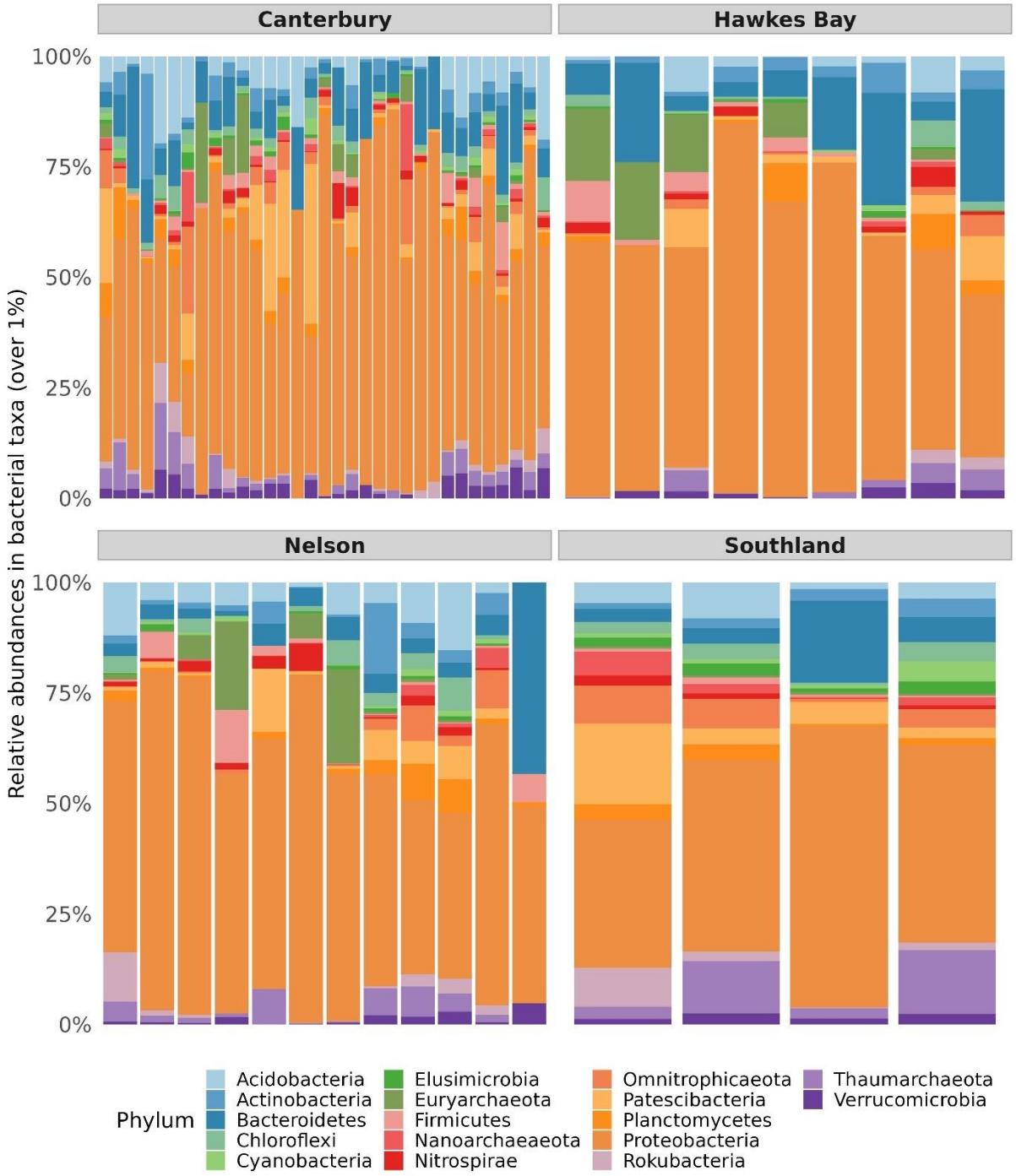


*Figure 2: Bacterial abundance over 1 % in wells across New Zealand*

## MODELLING

### PREDICTIONS WITH CART CLASSIFICATION TREE

The first node (root node) is the training data set, classifying the entire dataset (100 %) as below MAV (Figure 3). The internal node (decision-making node) splits into two sub nodes based on the threshold value of the relative abundance of Proteobacteria. When the relative abundance of Proteobacteria was greater than or equal to 60, the nitrate level was categorised as pristine with a probability of

90 %. When Proteobacteria had a relative abundance of less than 60, the groundwater was classified as below MAV with an 86 % certainty. Seventy-four percent of the groundwater samples fit into this split. This node divides further based on the best feature of the sub-group, and this final node (leaf node) holds the decision. Thus, if the relative abundance of Thaumarcheota was less than 1.9, then the groundwater was pristine (67 % accuracy). If the relative abundance of Thaumarcheota was not less than 1.9, the groundwater sample was below MAV (92 % accuracy). From the Thaumarcheota decision node, eight percent of the groundwater in this study was classified as pristine, and 67 % was below MAV.

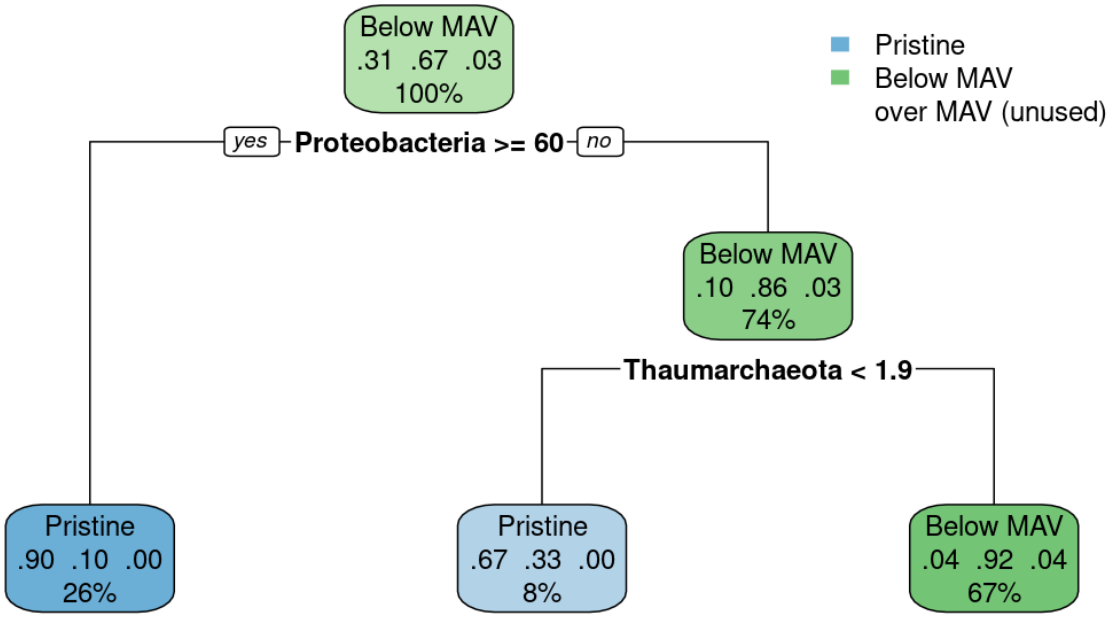


*Figure 3: Predictions with CART classification tree for WHO MAV nitrate levels*

Table 2. is a Prediction-Accuracy Table relating the CART model's observed and predicted outcome (confusion matrix). The CART model incorrectly classified two pristine predictions (they were below MAV) and two predictions of below MAV (one was pristine and was one above MAV).

*Table 2 Prediction-Accuracy Table for predicted and observed MAV nitrate levels for CART*

| | | Observed | | |
|---|---|---|---|---|
| | | Pristine | Below MAV | Above MAV |
| Prediction | Pristine | 11 | 2 | 0 |
| | Below MAV | 1 | 24 | 1 |
| | Above MAV | 0 | 0 | 0 |

A graphical representation of the CART classification model accuracy is shown in Figure 4. The points show the observed values from Table 2. The red points indicate an incorrect prediction, while the grey area is where the Below MAV samples should fall, and the white area is where the pristine samples should fall. This diagram collaborates Table 2. by showing that the two samples observed as Below MAV and the sample observed as Pristine were incorrectly predicted. The Above MAV point is not shown in this figure.

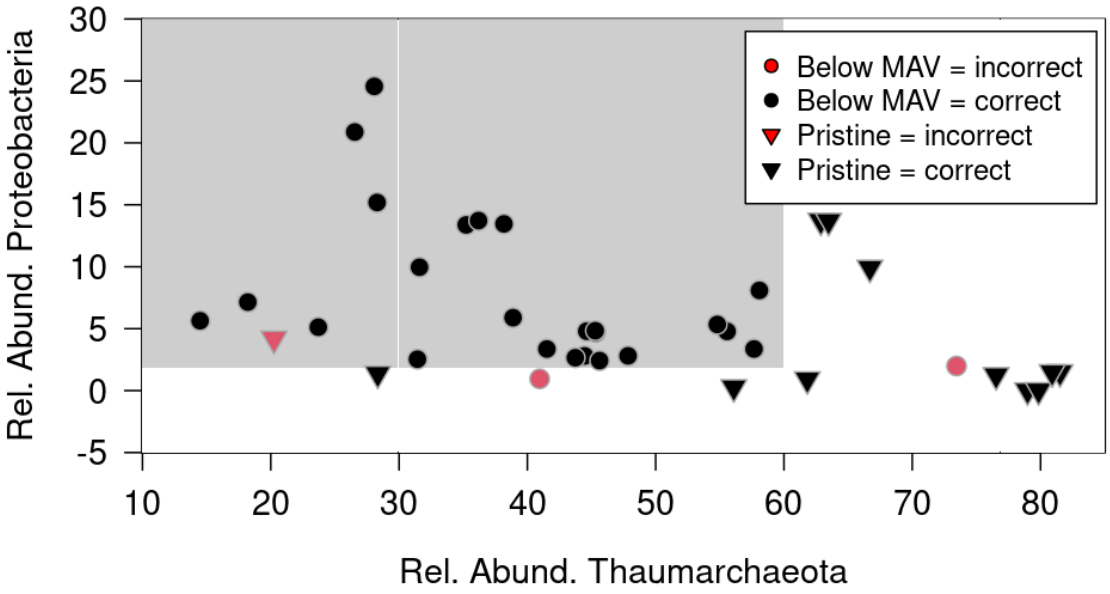


*Figure 4: Graphical representation of the CART classification showing the observed observations*

**PREDICTIONS WITH RANDOM FOREST**

Table 3. shows the confusion matrix for the Random Forest model. The model incorrectly predicted two samples as Pristine, three samples as Below MAV, and one sample as Above MAV. The Out of Box (OOB) estimate of the error for the Random Forest model was 12.82 %.

*Table 3 Prediction-Accuracy Table for predicted and observed MAV nitrate levels for Random Forest*

| | | Observed | | |
|---|---|---|---|---|
| | | Pristine | Below MAV | Above MAV |
| Prediction | Pristine | 10 | 2 | 0 |
| | Below MAV | 2 | 24 | 1 |

| | | | | |
|---|---|---|---|---|
| | Above MAV | 0 | 1 | 0 |

# DISCUSSION AND CONCLUSIONS

When the groundwater samples were classified according to Phyla (bacteria), both the CART classification and Random Forest models correctly predicted the nitrate classification most of the time. Random Forest is effectively a group of CARTs generating many random CART models while also taking a random selection of the variables provided to build each CART within the forest. Because Random Forest is based on the CART method, it will give similar outputs to CART, but with differences. Therefore, it is unsurprising that both the CART classification and Random Forest models gave similar results in this study.

While the initial model predictions are straightforward, they indicate that it may be possible to predict nitrate levels from microbial diversity data. The subset data used in this study had limited groundwater samples with above MAV nitrate levels. The next step in this study is to test the entire dataset in CART and Random Forest and test groundwater samples with a more extensive range of nitrate concentrations. We are also investigating predictions of nitrate levels on a higher level of microbial diversity, e.g., genera, beginning to look at the other critical predictors for overall groundwater health and test the robustness of using these predictors over a range of regions across New Zealand. The aim is to continue testing machine learning platforms to ultimately develop an in-line sensor that could detect changes in the abundance of selected microbes to predict contamination of groundwater resources.

**ACKNOWLEDGEMENTS**

We acknowledge the ESR Data Accelerator program for funding the research and Bridget Armstrong and Dr David Wood for being our mentors while undertaking this research.

**REFERENCES**

AKHTAR, N., SYAKIR ISHAK, M. I., BHAWANI, S. A. & UMAR, K. 2021. Various natural and anthropogenic factors responsible for water quality degradation: A review. Water, 13, 2660.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. & HOLMES, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods, 13, 581-583.

CAPORASO, J. G., LAUBER, C. L., WALTERS, W. A., BERG-LYONS, D., HUNTLEY, J., FIERER, N., OWENS, S. M., BETLEY, J., FRASER, L. & BAUER, M. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal, 6, 1621-1624.

GILPIN, B. J., WALKER, T., PAINE, S., SHERWOOD, J., MACKERETH, G., WOOD, T., HAMBLING, T., HEWISON, C., BROUNTS, A. & WILSON, M. 2020. A large

scale waterborne campylobacteriosis outbreak, Havelock North, New Zealand. Journal of Infection, 81, 390-395.

SALIS, R., BRUDER, A., PIGGOTT, J., SUMMERFIELD, T. & MATTHAEI, C. 2017. High-throughput amplicon sequencing and stream benthic bacteria: identifying the best taxonomic level for multiple-stressor research. Scientific reports, 7, 1-12.