CAWTHRON

REPORT NO. 3971

STOCKTAKE OF DATA QUALITY CODE AND METADATA USAGE IN REGIONAL COUNCIL WATER QUALITY DATASETS

World-class science
for a better future.

STOCKTAKE OF DATA QUALITY CODE AND METADATA USAGE IN REGIONAL COUNCIL WATER QUALITY DATASETS

JESSICA SCHATTSCHNEIDER, ROGER YOUNG

Prepared for Ministry for the Environment

CAWTHRON INSTITUTE
98 Halifax Street East, Nelson 7010 | Private Bag 2, Nelson 7042 | New Zealand
Ph. +64 3 548 2319 | Fax. +64 3 546 9464
www.cawthron.org.nz

REVIEWED BY:
Laura Kelly

APPROVED FOR RELEASE BY:
Joanne Clapcott

ISSUE DATE: 18 September 2023

RECOMMENDED CITATION: Schattschneider J, Young R. 2023. Stocktake of data quality code and metadata usage in regional council water quality datasets. Nelson: Cawthron Institute. Cawthron Report 3971. Prepared for Ministry for the Environment.

DISCLAIMER: While Cawthron Institute (Cawthron) has used all reasonable endeavours to ensure that the information contained in this document is accurate, Cawthron does not give any express or implied warranty as to the completeness of the information contained herein, or that it will be suitable for any purpose(s) other than those specifically contemplated during the project or agreed by Cawthron and the client.

EXECUTIVE SUMMARY

Reliable data on the health of our waterways is critical for adaptive management. One component of efforts in Aotearoa New Zealand to improve the reliability and consistency of environmental data has been the introduction of quality control (QC) codes. A QC code schema was developed in 2013 as part of the National Environment Monitoring Standards (NEMS) initiative. Councils and other data providers have been encouraged to use the QC code system to gauge the accuracy of individual data records, but the level of adoption of QC codes for waterway health data has been unclear. Similarly, the use of metadata connected with data records has been encouraged, but the level of adoption of good metadata practices is unknown. In this report we describe the use of QC codes and metadata across three data streams (referred to in this report as modules): rivers (water quality parameters for rivers), macroinvertebrates (for rivers) and lakes (a mix of water quality and biological data). The analysis is based on data over the period 2004–21 that was sourced from council data servers between June to September 2022 as part of the annual water quality update for Land, Air, Water Aotearoa (LAWA).

Based on the waterway health data that were sourced, 11 of 16 councils have adopted QC codes to some extent. QC code use is most prevalent for the rivers module, where approximately half of river data records have been assigned a QC code, whereas adoption of QC codes is lower for macroinvertebrate (25%) and lakes (13%) data. For the data that do have QC codes, about half use internal council QC codes rather than the NEMS coding system. Five councils have applied QC coding to almost all of their rivers data, but adoption of QC coding is more limited for other modules and for other councils.

There was no clear pattern in the adoption of QC codes after the release of the NEMS QC coding schema in 2013. Some councils were either using their own QC coding systems prior to this, or have retrospectively applied QC codes to data collected prior to 2013. Some councils have applied QC coding of their data in bursts of activity, followed by periods where QC codes have not been applied.

We found evidence for differences in the statistical distribution of values among data classified as Good quality, Synthetic, Uncoded and Poor quality. This emphasises the importance of adopting QC coding, as analyses of data of unknown quality may result in a different outcome to analyses focusing only on Good quality data.

Based on the information that was sourced, 15 of 16 councils have recorded metadata alongside their waterway health data. These metadata predominantly include administrative information (e.g. project name, sample ID, field technician name), weather conditions at the time of sampling, field instrument details, and laboratory sample analysis methods (e.g. laboratory name, analysis method, detection limit). We identified a substantial number of metadata variables for rivers (> 195), macroinvertebrates (86) and lakes (178), indicating a significant lack of consistency in how councils capture and manage this information. This

inconsistency highlights the need for standardised approaches to ensure greater uniformity in metadata practices among councils.

We acknowledge certain limitations in our work that should be considered when analysing the results presented here. First, many councils are in an ongoing process of enhancing their data publishing methods and upgrading their data servers. Consequently, as this report relies on data obtained from council data servers during the 2022 LAWA freshwater refresh, any subsequent data improvements made by the councils on their servers since late 2022 are not reflected in our analyses. Furthermore, the extraction of QC and metadata from council servers posed challenges. While we received feedback from some councils on locating this information on their servers, complete instructions on accessing such information, especially for metadata extraction from councils not using the Hilltop server, were not provided. Therefore, we acknowledge that councils may record and store QC and metadata information in other data sources to which we did not have access, and as a result, these sources were not included in our study.

Our analyses provide a snapshot view of the adoption of QC codes and use of metadata, and highlight the current high level of variability in adoption among councils and across different datasets. This report serves as a potential catalyst for increased adoption among councils that are just starting on this journey. Key insights from our study include:

1. Diverse levels of QC code adoption are applied across modules and councils.
2. QC codes can be applied retrospectively.
3. There is widespread inconsistency in metadata capture and management, both within and across councils.
4. A combination of NEMS and internal codes are in use.

To enhance the availability of consistent datasets for efficient waterway management, we recommend that the following steps are taken:

1. Investigate the primary challenges in adopting NEMS codes at a council and module level.
2. Further explore the processes behind retrospective QC code application.
3. Develop guidelines and standards for the use and interpretation of child codes to help inform decisions on whether data should be excluded from subsequent analyses.
4. Establish clear and comprehensive metadata standards to ensure consistency within and across councils.
5. Extract lessons from councils proficient in QC code adoption to assist others.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

GLOSSARY

AC	Auckland Council
API	Application programming interface
ASPM	Average score per metric
BDISC	Visual clarity
BOPRC	Bay of Plenty Regional Council
CYANOTOT	Total cyanobacteria biovolume
CYANOTOX	Potentially toxic cyanobacteria biovolume
DIN	Dissolved inorganic nitrogen
DRP	Dissolved reactive phosphorus
ECAN	Environment Canterbury
ECOLI	*Escherichia coli*
ES	Environment Southland
GDC	Gisborne District Council
GWRC	Greater Wellington Regional Council
HBRC	Hawke’s Bay Regional Council
HRC	Horizons Regional Council
MCI	Macroinvertebrate Community Index
MDC	Marlborough District Council
NCC	Nelson City Council
NEMS	National Environmental Monitoring Standards
NH4N	Ammoniacal nitrogen
NO3N	Nitrate nitrogen
NRC	Northland Regional Council
ORC	Otago Regional Council
QMCI	Quantitative Macroinvertebrate Community Index
SECCHI	Secchi disc depth
TDC	Tasman District Council
TN	Total nitrogen
TON	Total oxidised nitrogen
TP	Total phosphorus
TRC	Taranaki Regional Council
TURB	Turbidity – Nephelometric Turbidity Units
TURBFNU	Turbidity – Formazin Turbidity Units
WCRC	West Coast Regional Council
WRC	Waikato Regional Council
% EPT Taxa	Percentage of Ephemeroptera, Trichoptera and Plecoptera taxa

1. INTRODUCTION

Reliable water quality data are vital to the management of freshwater resources. Quality assurance / quality control (QA / QC) measures are used to ensure that the data collected are accurate and that monitoring procedures adhere to established standards. QA involves a comprehensive plan for maintaining quality throughout a programme, while QC involves determining the validity of specific sampling and analytical procedures to ensure that the information collected is accurate, precise and properly recorded (DES 2018).

QC code schemas can help to minimise errors in water quality data and provide confidence in data analysis. Metadata, which include details about the sampling process, laboratory analysis and any other factors that may influence the quality of the data, are also an important aspect of water quality monitoring. Capturing and storing metadata alongside the data allows for greater understanding and interpretation of the results, as well as facilitating sharing and reuse of data.

The National Environmental Monitoring Standards (NEMS) programme has aimed to improve consistency in environmental monitoring practices by relevant agencies across Aotearoa New Zealand. The NEMS steering group has facilitated the production of a wide range of standards to improve consistency in the way data are collected, processed and analysed.[1] It is recommended that these standards are adopted by all agencies throughout Aotearoa New Zealand that are involved in environmental monitoring.

The national quality code schema was developed as part of the NEMS programme and was first issued in June 2013 (NEMS 2013). This instrument aims to allow end-users to consistently use and / or review environmental data sourced from multiple organisations, and to provide a better understanding of the data collection methodologies and data limitations. NEMS codes use a numeric index from 0 to 600, with higher numbers generally meaning better-quality data. In addition to the parent quality codes (Figure 1), the NEMS also allows an expanded set of supplementary quality codes, or child codes. Child codes are currently allocated in-house, which gives councils the ability to expand the QC series. For example, parent code QC 200 could be expanded to child code 210 and 250 to differentiate between data that are of known quality and data that are currently non-verified (Figure 2).

[1] See www.nems.org.nz

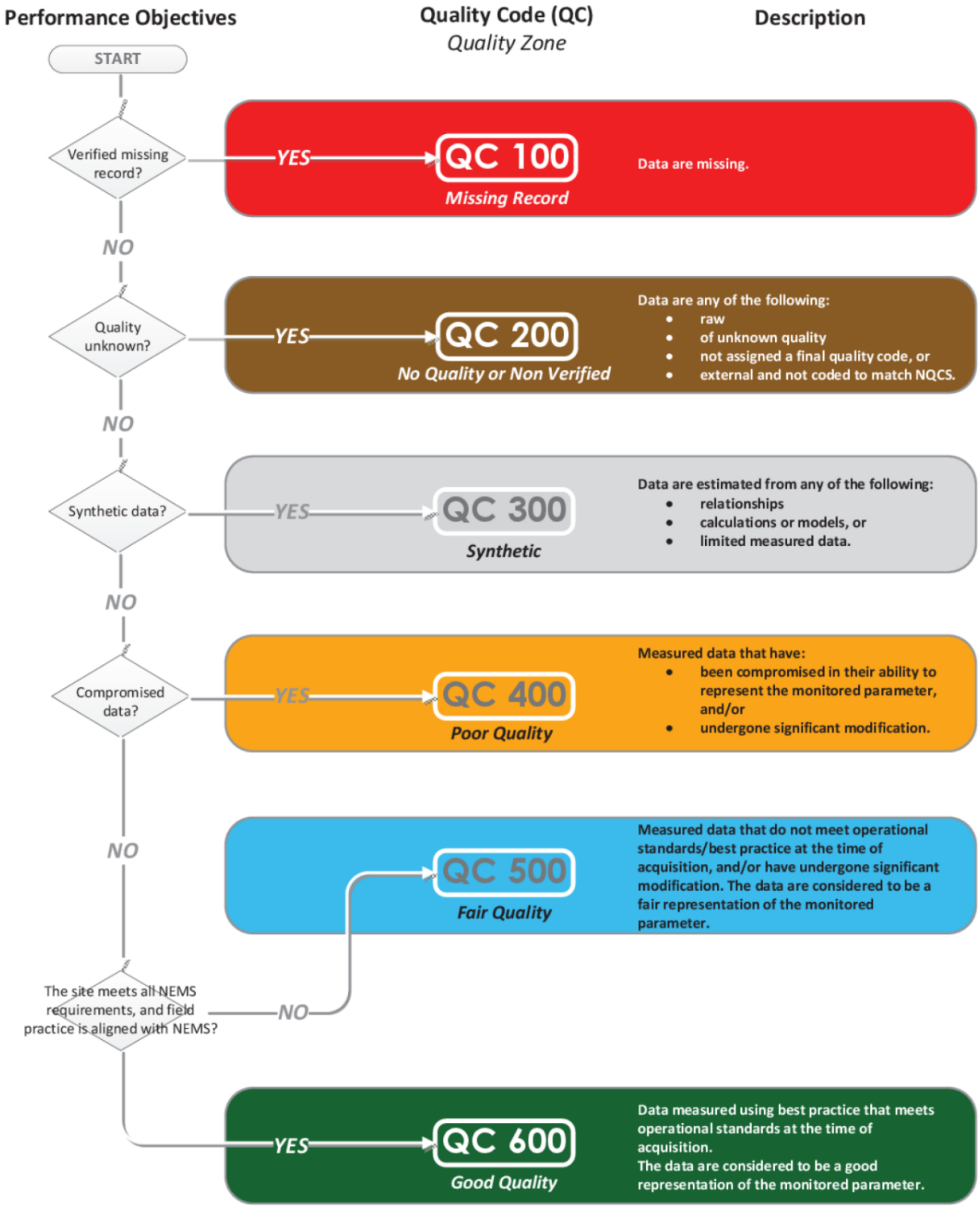
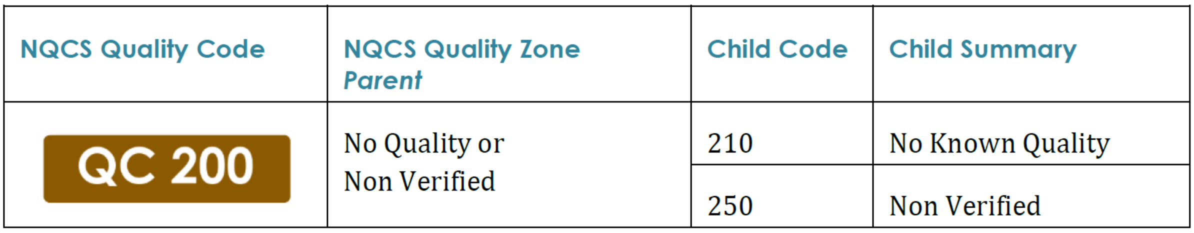

Performance Objectives
Quality Code (QC)
Quality Zone
Description
START
Verified missing record?
YES
QC 100
Missing Record
Data are missing.
NO
Quality unknown?
YES
QC 200
No Quality or Non Verified
Data are any of the following:
• raw
• of unknown quality
• not assigned a final quality code, or
• external and not coded to match NQCS.
NO
Synthetic data?
YES
QC 300
Synthetic
Data are estimated from any of the following:
• relationships
• calculations or models, or
• limited measured data.
NO
Compromised data?
YES
QC 400
Poor Quality
Measured data that have:
• been compromised in their ability to represent the monitored parameter, and/or
• undergone significant modification.
NO
QC 500
Fair Quality
Measured data that do not meet operational standards/best practice at the time of acquisition, and/or have undergone significant modification. The data are considered to be a fair representation of the monitored parameter.
The site meets all NEMS requirements, and field practice is aligned with NEMS?
NO
YES
QC 600
Good Quality
Data measured using best practice that meets operational standards at the time of acquisition.
The data are considered to be a good representation of the monitored parameter.

Figure 1. Generic quality flowchart showing the meaning of each QC code. Source: Reprinted from NEMS (2016).



NQCS Quality Code	NQCS Quality Zone *Parent*	Child Code	Child Summary
QC 200	No Quality or Non Verified	210	No Known Quality
		250	Non Verified

Figure 2. An example of a parent QC code and associated child QC codes used to provide more resolution to the QC information. Source: Reprinted from NEMS (2016).

The NEMS reports on discrete river water quality data (NEMS 2019a), discrete lake water quality data (NEMS 2019b), and macroinvertebrate data from rivers and streams (NEMS 2022) provide details on how councils should implement QA / QC procedures and recommend how QC codes and metadata should be assigned and stored with individual measurements for each variable. All three reports state that data shall be quality coded in accordance with the NEMS quality code schema (NEMS 2016), and that metadata shall be stored together in a time-series data server, linked with a single date and time, to ensure standardisation of datasets, enabling the comparison of data within regions, across regions and nationally.

The NEMS states that each measurement shall be stored with the following information:

- its associated measurement date, time and units
- field instrumentation (make, model and number) or laboratory name, location and test method
- clear reference to its associated form (dissolved, total, reactive, etc.), where applicable
- all relevant visit-related metadata, including the name(s) of personnel conducting field measurements and sampling
- relevant laboratory comments, where applicable
- processing laboratory name and location
- processing method and laboratory staff processing identification number.

The purpose of this report is to investigate the use of quality codes and metadata in water quality data collected by the 16 local government councils of Aotearoa New Zealand. By doing this, we aim to identify opportunities to enhance data consistency and improve subsequent analyses. To achieve this, we analysed three types of data, which we refer to as ‘modules’ throughout the report: river water quality data (referred to as the ‘rivers’ module), lake water quality data (referred to as the ‘lakes’ module) and macroinvertebrate metrics (referred to as the ‘macro’ module). In this report, individual types of metrics or variables from these modules are referred to as ‘parameters’, such as chlorophyll-*a* (CHLA), total nitrogen (TN) and turbidity (TURB). A single observation or measurement is referred to as a ‘record’.

To identify opportunities to enhance data consistency, we examined the frequency of use of different codes associated with various parameters across different councils. We also assessed the adoption of internal QC code schemas and compared them to the NEMS coding schema. It is important to note that our study primarily focused on the adoption of the parent codes from the NEMS coding schema. However, we used council feedback provided during the data compilation process to identify and define the meaning of some child codes, and also interpreted the meaning of certain internal codes. Any other codes for which feedback was not provided by the agency using them were classified as internal codes, and we did not attempt to translate their meaning or relate them to NEMS codes. Furthermore, we investigated the availability of metadata for each council to supplement our analysis.

The body of this report is structured into nine sections, with this introduction forming the first section. Section 2 details the process of how the data for analysis were obtained and processed. Section 3 focuses on presenting general summaries of how QC codes and QC schemas are being used in different councils, modules and individual parameters. In Section 4 we investigate the adoption of QC codes throughout the years for each council and the possible shift of coding schemas over time. In Section 5 we examine whether data with different codes exhibit differences in descriptive statistics. Section 6 presents a general overview of the metadata stored in the dataset analysed in this project. And finally, in Section 7 we wrap up our work, highlighting the main findings, next steps and recommendations for data improvements.

2. DATA COMPILATION AND METHODOLOGY

All the water quality data used in this report were collated as part of the annual Land, Air, Water Aotearoa (LAWA) 2022 freshwater data refresh and published by LAWA in September 2022.[2] At the time of writing this report, LAWA contains data from councils from 2004 until 2021. The URL addresses used in the LAWA 2022 freshwater data refresh to pull the dataset analysed here are listed in Appendix 2. Analyses were conducted using R (R Core Team 2022) and RStudio (RStudio Team 2022).

2.1. QC codes

The focus of the LAWA 2022 refresh was to collate QC code information alongside water quality data. On a weekly basis from June 2022 to September 2022, Cawthron Institute (Cawthron) accessed water quality data from council servers. Weekly reports were then generated summarising the data into tables that presented the unique QC codes used by each council and the number of records associated with these codes. Councils were advised to review the weekly QC code summaries to confirm that Cawthron was successfully extracting all the available QC codes and that all their QC information had been added to the data server at the time the data were sourced. As part of the annual LAWA refresh, councils' replies were addressed and used to create the final version of the QC code dataset used in this report.

Based on QC information from councils that provided feedback, we were able to translate some internal council codes to the NEMS 'parent' and 'child' schema before starting to analyse the data (Table 1). If councils had internal schemas matching the NEMS schema codes, these were classified as NEMS codes. Overall, the analyses related to QC codes in this report relied heavily on how councils recorded QC codes in their server and their level of engagement in the data-pull process during the LAWA 2022 refresh.

[2] https://www.lawa.org.nz

Table 1. Internal codes and NEMS child codes translated to parent NEMS codes according to feedback from councils. This translation was conducted before analysing the dataset and generating the results presented in this report. AC – Auckland Council, ES – Environment Southland, HBRC – Hawke's Bay Regional Council, MDC – Marlborough District Council.

Council	Original code	NEMS translation	Definition
AC	0, 16384, 10	600	Good quality
AC	9, 16393	610	NEMS child code – 600 parent
AC	21	500	NEMS 500 – Fair quality
AC	25, 16409	560	NEMS child code – 500 parent
AC	26	550	NEMS child code – 500 parent
AC	35, 16419	460	NEMS child code – 400 parent
AC	36	450	NEMS child code – 400 parent
AC	39, 16423	543	NEMS child code – 500 parent
AC	41,61,42,151,16425	400	Poor quality
AC	51, 8243	200	No quality
AC	61	100	Incorrect or missing
ES	403,404,406	400	Poor quality
HBRC	40	400	Poor quality
MDC	450	400	Poor quality

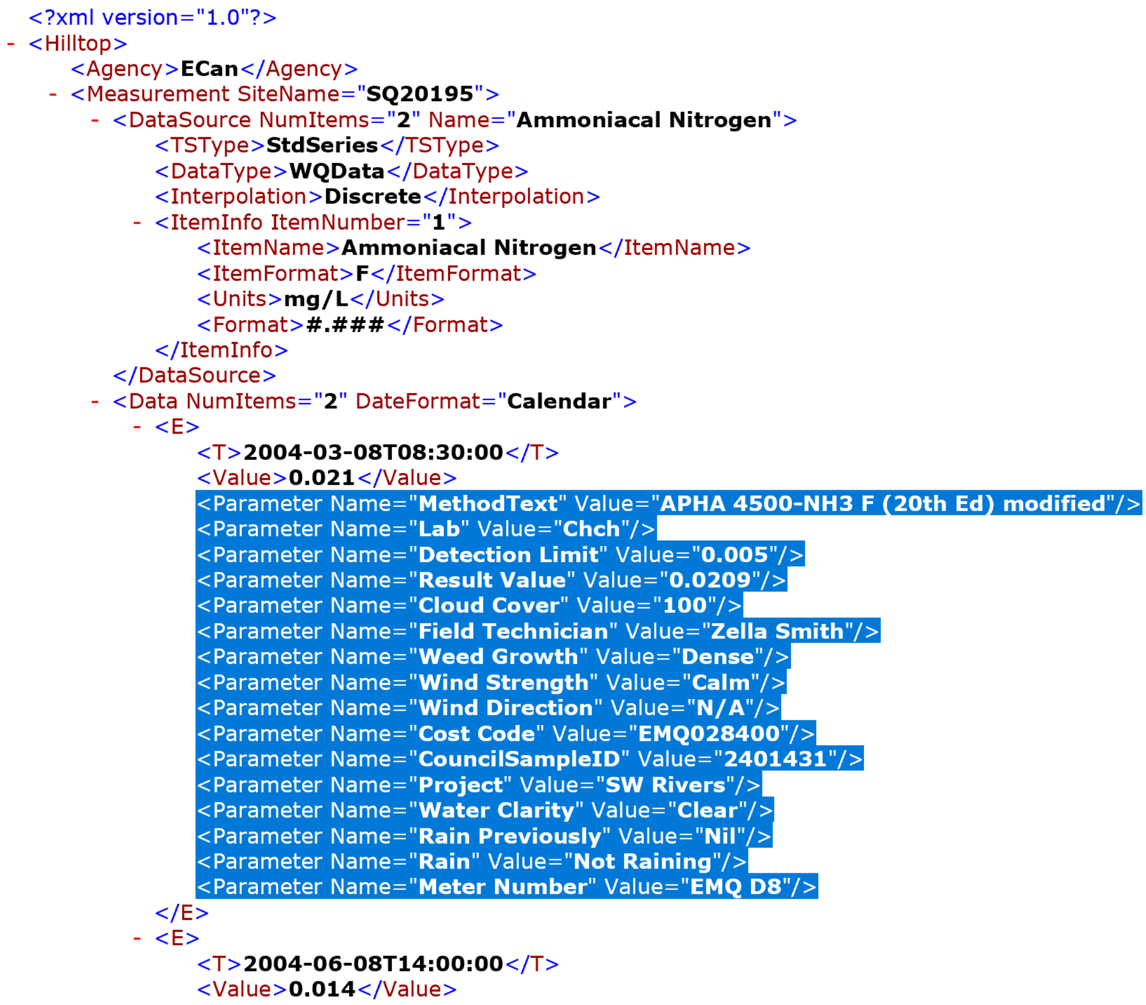
Prior to conducting our analyses of the water quality data, we also took into consideration censored data, which are data records that are either below a minimum detection limit (left-censored) or above a maximum detection limit (right-censored) set by the laboratory. To address the impact of censored data on our results, we followed the approach adopted by LAWA, where left-censored values were halved, and right-censored values were multiplied by 1.1. We also removed records lower than zero, as measured values cannot be negative.

To explore whether there were differences in the distribution of water quality data classified according to different QC codes, we plotted the data using box and whisker plots. We continued this analysis for a select few parameters, further exploring the potential for differences using linear regression models in R. For this, we selected one parameter from each module, focusing on parameters with more than 30 records in each QC group to ensure robustness. We used a generalised linear mixed model, with the value of the indicator as the response variable and the QC group as the explanatory variable. To account for non-independence of the response variable, we used site as the random effect in the models. The primary objective of these models was to determine possible differences in data distribution between QC groups, while also accounting for the categorical effect of measurements obtained from different sites. This was done because the values of the parameter may be influenced by the specific site where the data were collected. This analysis aimed to investigate whether any significant differences existed in the distribution of values across different QC groups, as such disparities could lead to varied results in subsequent analyses. However, it is important to note that our focus was on identifying the presence of differences rather than determining their specific direction. Initially, we also attempted to include council as a random effect, because each council follows its own distinct data collection and QC procedures, which may introduce variations in the measured values. In addition, we tried to include season as a response variable, as values can vary greatly throughout the year. However, the inclusion of these elements resulted in convergence issues and numerical instability, indicating challenges in model fitting. As a result, we decided to focus solely on site as a random effect in our analysis.

2.2. Metadata

For the metadata analysis, we used the raw XML (eXtensible Markup Language) files obtained from councils' data servers during the LAWA 2022 refresh. XML files are a text-based file format used for storing and exchanging structured data. They consist of a hierarchical structure made up of elements, attributes and text content.

Figure 3 provides an example of the raw XML file retrieved from Environment Canterbury's (ECAN) data server, representing a time series of ammoniacal nitrogen for a specific site. In the highlighted portion of this figure, the metadata for the record from 2004-03-08T08:30:00 (8 March 2004, 8.30am) can be observed. The 'Parameter Name' XML attribute indicates the name of the metadata variable, while the 'Value' attribute stores the corresponding value of the metadata.



```
  <?xml version="1.0"?>
- <Hilltop>
      <Agency>ECan</Agency>
    - <Measurement SiteName="SQ20195">
      - <DataSource NumItems="2" Name="Ammoniacal Nitrogen">
            <TSType>StdSeries</TSType>
            <DataType>WQData</DataType>
            <Interpolation>Discrete</Interpolation>
          - <ItemInfo ItemNumber="1">
                <ItemName>Ammoniacal Nitrogen</ItemName>
                <ItemFormat>F</ItemFormat>
                <Units>mg/L</Units>
                <Format>#.###</Format>
            </ItemInfo>
        </DataSource>
      - <Data NumItems="2" DateFormat="Calendar">
          - <E>
                <T>2004-03-08T08:30:00</T>
                <Value>0.021</Value>
                <Parameter Name="MethodText" Value="APHA 4500-NH3 F (20th Ed) modified"/>
                <Parameter Name="Lab" Value="Chch"/>
                <Parameter Name="Detection Limit" Value="0.005"/>
                <Parameter Name="Result Value" Value="0.0209"/>
                <Parameter Name="Cloud Cover" Value="100"/>
                <Parameter Name="Field Technician" Value="Zella Smith"/>
                <Parameter Name="Weed Growth" Value="Dense"/>
                <Parameter Name="Wind Strength" Value="Calm"/>
                <Parameter Name="Wind Direction" Value="N/A"/>
                <Parameter Name="Cost Code" Value="EMQ028400"/>
                <Parameter Name="CouncilSampleID" Value="2401431"/>
                <Parameter Name="Project" Value="SW Rivers"/>
                <Parameter Name="Water Clarity" Value="Clear"/>
                <Parameter Name="Rain Previously" Value="Nil"/>
                <Parameter Name="Rain" Value="Not Raining"/>
                <Parameter Name="Meter Number" Value="EMQ D8"/>
            </E>
          - <E>
                <T>2004-06-08T14:00:00</T>
                <Value>0.014</Value>
```

Figure 3. Example of XML data retrieved by Hilltop servers. The selected lines are the portion of information (metadata) we aim to extract. Note that the metadata variable is called the ‘Parameter Name’ attribute and the metadata value is available in the ‘Value’ attribute of this file.

To analyse these data, we extracted the metadata (when available) from the XML files and combined them into a table. We did this by developing R script specifically to perform parsing[3] operations on the XML files obtained from the councils’ servers.

It is worth noting here that while we had accompanying background information from councils relating to the QC component of this report (Section 2.1), it was outside the scope of this study to request that councils provide similar background information relating to their metadata. Therefore, we analysed metadata only from XMLs collected during the LAWA 2022 refresh using the documentation (when) available for their data

[3] Parsing refers to the process of analysing the structure and content of a file or document to extract specific information or convert it into a structured format. In our study, this involved interpreting the hierarchical structure of the XML files, identifying the relevant elements and attributes containing the metadata, and transforming them into a table format.

servers to create a generic script to automate the metadata extraction. We also analysed metadata only from councils that provided their data via a data server connection. This means that metadata from councils that share data in spreadsheet files, for example, were not evaluated.[4]

As part of our investigation into the level of standardisation and consistency in how metadata is stored on councils' servers, we purposely refrained from conducting any common text cleaning processes on the original data. This means that we did not perform actions such as identifying synonyms, or removing trailing spaces from the metadata as it existed in its original form. However, for each council, we did remove all unique metadata variables that appeared less than 10 times in the dataset before performing the analyses presented in this report and for some descriptive analyses we converted all words to lower case.

It is recommended in NEMS (2016) that water quality data and visit metadata should be stored together in a time-series data server linked with a single date and time. Hence, we only pulled metadata information available for each individual record in the XML files. Any batch system to assign metadata from one record to the following ones was not captured here.

[4] This report does not include data on macroinvertebrates from sites monitored by Auckland Council, or macroinvertebrate and water quality data from sites monitored by the National Institute of Water and Atmospheric Research (NIWA). These datasets were provided by spreadsheets during the LAWA Refresh 2022 and were therefore not included or analysed in this report.

3. ADOPTION OF QUALITY CONTROL CODES

There were substantial differences in the adoption rates of QC codes across the three different modules (Table 2). The river water quality module (rivers) had the highest adoption rate, with half of the data being linked with QC codes. This was followed by the macroinvertebrates module (macro), with 25% of data linked with QC codes, while only 13% of the data from the lake water quality module (lakes) had linked QC codes. It is worth noting that just about half of the codes used are from the NEMS schema.

Table 2. Proportion of records with QC codes per module. The last column indicates the proportion of codes used from the NEMS schema.

	Total number of records	QC coded records (%)	QC coded records with NEMS code (%)
Rivers	709,298	49.92	28.75
Macroinvertebrates	18,146	25.02	16.19
Lakes	12,271	13.05	7.79

The adoption of QC codes and QC schemas varied significantly among councils, with only 11 out of 16 councils having implemented QC codes to some extent (Figure 4). The code schema used by the 11 councils that did implement codes varied between them. Horizons Regional Council (HRC) and Greater Wellington Regional Council (GWRC) used a combination of internal and NEMS codes (predominantly using the NEMS schema), while six councils – including Auckland Council (AC), ECAN, Environment Southland (ES), Gisborne District Council (GDC), Marlborough District Council (MDC) and Nelson City Council (NCC) – used only NEMS codes (or internal codes that could be translated to NEMS codes; Table 1). West Coast Regional Council (WCRC) also used only codes from the NEMS schema, but for a very small portion of their rivers data (0.05%). Only two councils, Hawke's Bay Regional Council (HBRC) and Waikato Regional Council (WRC), used internal codes exclusively. The five remaining councils surveyed are yet to implement some form of QC coding; they are Bay of Plenty Regional Council (BOPRC), Northland Regional Council (NRC), Otago Regional Council (ORC), Tasman District Council (TDC) and Taranaki Regional Council (TRC).

The breakdown of the number of records associated to each individual code and their corresponding code schema is presented in Table 3.

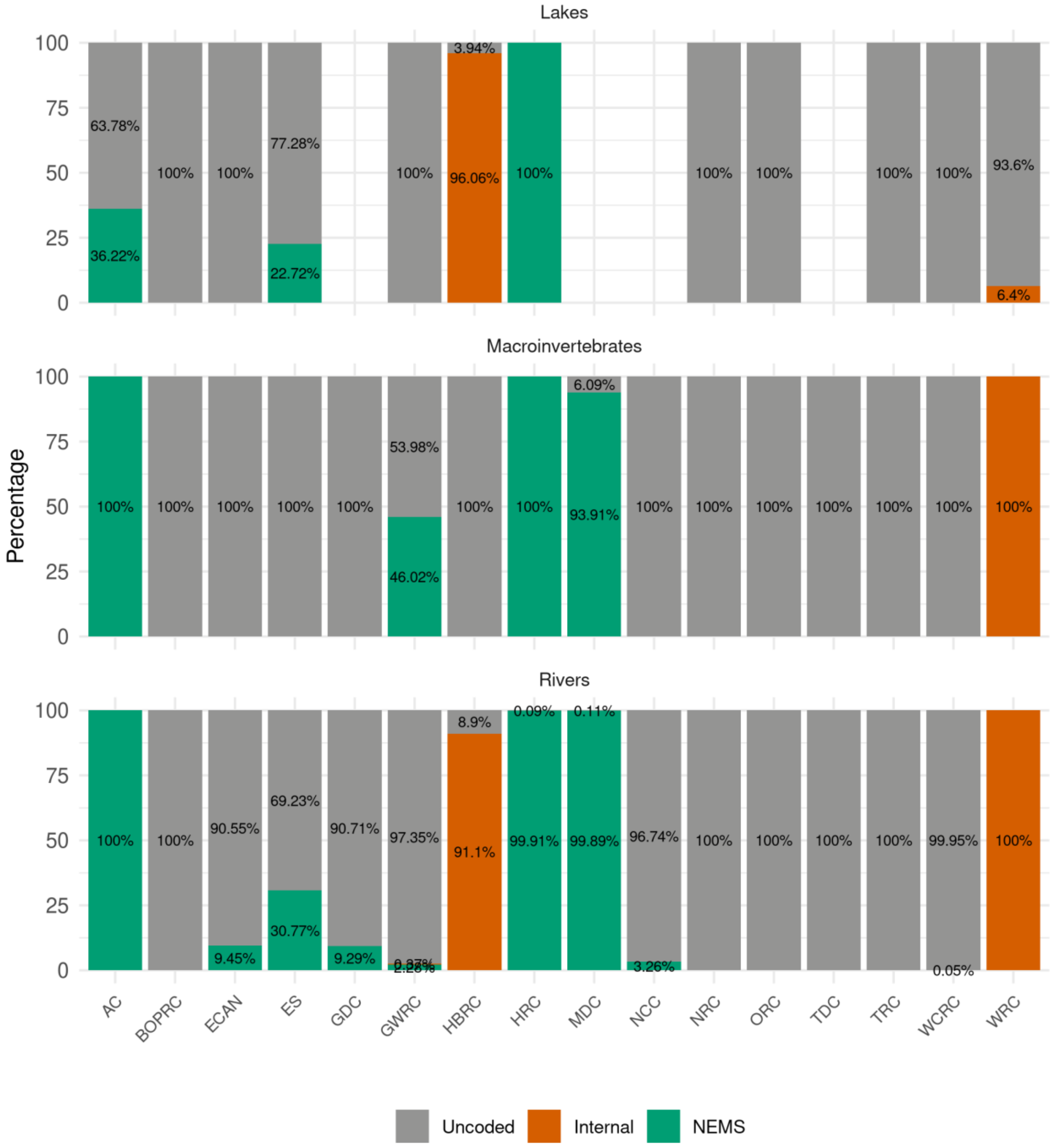


Figure 4. Proportion of records with QC codes per council associated with the rivers, lakes and macro modules. For an explanation of abbreviations and terms, see Glossary.

Table 3. Number of records associated with each unique NEMS or internal council QC code for each council. '—' indicates that no records were associated to a given code. The results are separated per module and the code schema that each QC code pertains to is specified in the column 'Code schema'. For an explanation of abbreviations and terms, see Glossary.

	Code schema	AC	BOPRC	ECAN	ES	GDC	GWRC	HBRC	HRC	MDC	NCC	NRC	ORC	TDC	TRC	WCRC	WRC
Rivers																	
100	NEMS	3	—	—	—	—	—	—	113	—	—	—	—	—	—	—	—
200	NEMS	678	—	—	29988	1301	—	—	3872	18	—	—	—	—	—	17	—
300	NEMS	—	—	16375	—	2542	—	—	24477	—	—	—	—	—	—	—	—
400	NEMS	901	—	—	745	2	521	—	—	1	—	—	—	—	—	—	—
450	NEMS	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
460	NEMS	1938	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
500	NEMS	12	—	—	4	610	356	—	2891	5	96	—	—	—	—	—	—
543	NEMS	66	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
550	NEMS	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
560	NEMS	811	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
600	NEMS	5720	—	—	7941	214	1353	—	195869	44485	790	—	—	—	—	—	—
610	NEMS	63834	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	Internal	—	—	—	—	—	360	—	—	—	—	—	—	—	—	—	—
10	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	28676
50	Internal	—	—	—	—	—	—	6572	—	—	—	—	—	—	—	—	—
60	Internal	—	—	—	—	—	—	100111	—	—	—	—	—	—	—	—	—
90	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	19
210	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	163981
219	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
222	Internal	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—
225	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	65
228	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	894
231	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	62
NA	Uncoded	—	54365	156987	87025	45568	95321	10426	196	51	26283	65673	82043	22153	29333	36233	—
Macroinvertebrates																	
200	NEMS	—	—	—	—	—	1637	—	4923	—	—	—	—	—	—	—	—
600	NEMS	3750	—	—	—	—	—	—	—	1433	—	—	—	—	—	—	—
210	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	6392
254	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	11
NA	Uncoded	—	6712	13625	5700	2425	1920	3438	—	93	1180	2856	2286	681	9764	3690	—
Lakes																	
200	NEMS	—	—	—	2430	—	—	—	226	—	—	—	—	—	—	—	—
400	NEMS	12	—	—	85	—	—	—	—	—	—	—	—	—	—	—	—
450	NEMS	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
460	NEMS	30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
500	NEMS	2	—	—	—	—	—	—	35	—	—	—	—	—	—	—	—
560	NEMS	80	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
600	NEMS	577	—	—	194	—	—	—	3031	—	—	—	—	—	—	—	—
610	NEMS	615	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
10	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	254
20	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	75
50	Internal	—	—	—	—	—	—	757	—	—	—	—	—	—	—	—	—
60	Internal	—	—	—	—	—	—	3165	—	—	—	—	—	—	—	—	—
210	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	289
213	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	397
231	Internal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	12
NA	Uncoded	2326	12601	19765	9216	—	2308	161	—	—	—	9358	8133	—	914	1932	15032

We further examined the use of different QC code schemas by councils across different water quality parameters. The following figures in this section are based on the values presented in Appendix 1.

In the rivers module, AC, HRC and MDC were found to employ the NEMS schema to QC code all records for the parameters they monitor within this module (Figure 5). Similarly, ECAN has adopted the NEMS schema to QC code all their dissolved inorganic nitrogen (DIN) records. GDC and ES also use the NEMS schema to QC code a portion of their data across different parameters.

Conversely, HBRC and WRC exclusively use internal codes for QC coding purposes, as highlighted in the previous results, with the exception of DIN data for HBRC, which is Uncoded.

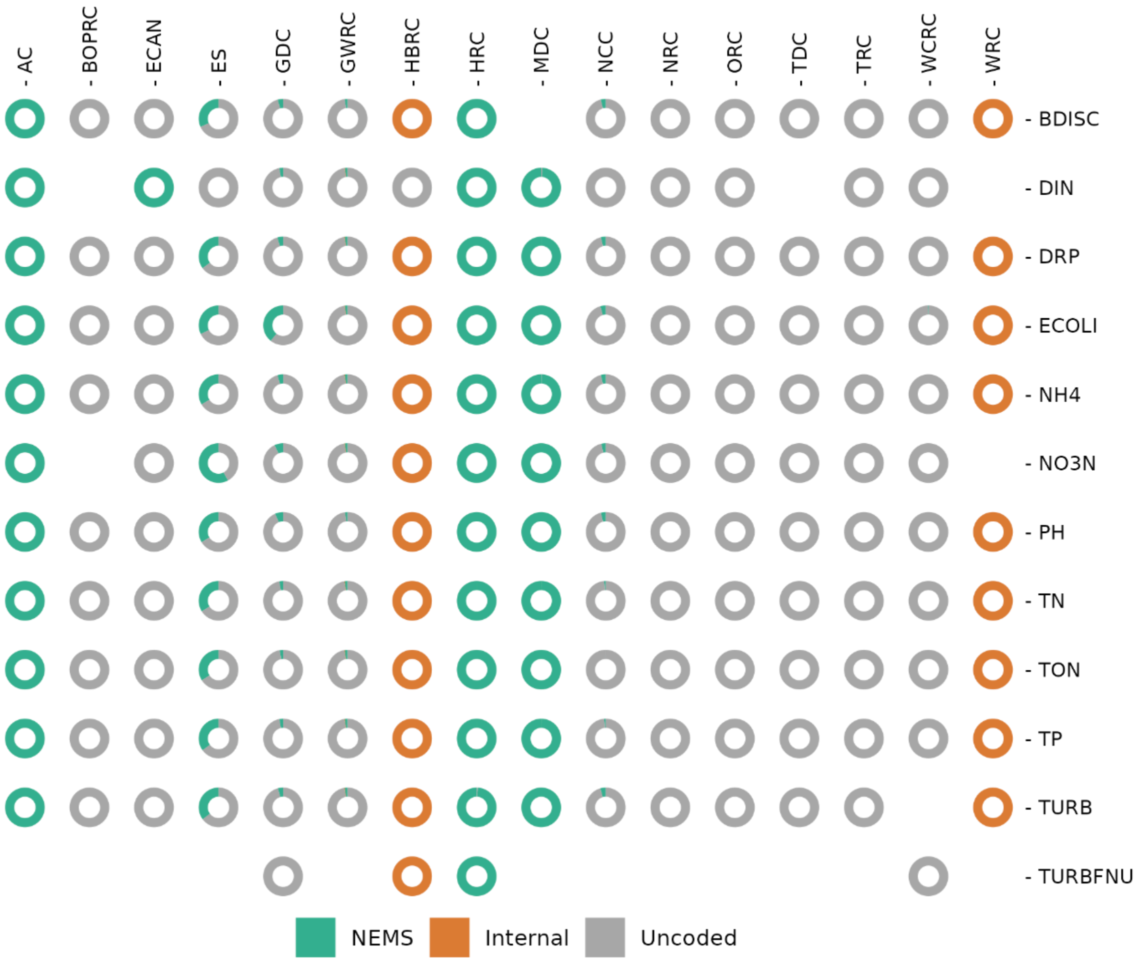


Figure 5. Proportion of records using NEMS QC codes (shown in green) and internal QC codes (shown in orange) for each parameter of the rivers module per council. The proportion of Uncoded records is represented in grey. Parameters that are not measured by councils are presented as missing cells. For an explanation of abbreviations and terms, see Glossary.

In the case of the macroinvertebrates module, councils use NEMS codes across a substantial portion of the QC coded data (Figure 6). AC and HRC QC code all of their macroinvertebrate data, while MDC QC codes almost all of its data. Interestingly, AC and MDC use only NEMS QC code 600, which signifies Good quality data, whereas HRC relies solely on NEMS QC code 200, indicating Uncoded data (Table 1, Table 3).

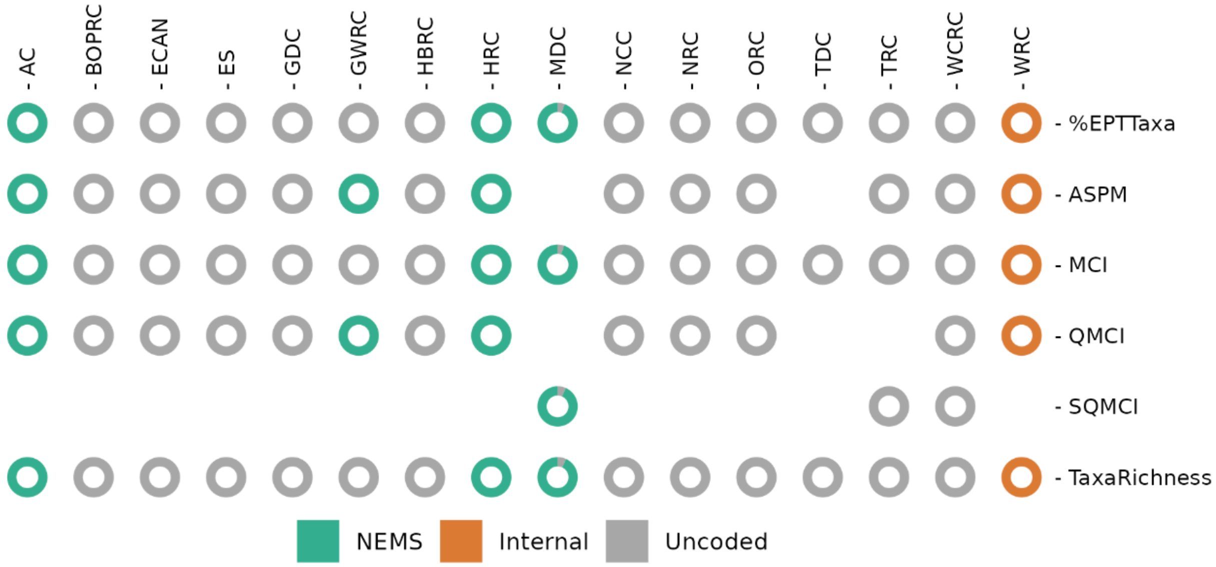


Figure 6. Proportion of records using NEMS QC codes (shown in green) and internal QC codes (shown in orange) for each parameter of the macroinvertebrates module per council. The proportion of Uncoded records is represented in grey. Parameters that are not measured by councils are presented as missing cells. For an explanation of abbreviations and terms, see Glossary.

Among the analysed modules, lakes data exhibited the lowest adoption of QC codes, which is reflected in the analysis of lakes parameters (Figure 7). Once again, HRC emerges as one of the councils with the highest QC code implementation, applying QC codes to all their lakes data using the NEMS code schema. However, the majority of these data are associated with only two codes, 600 and 200, representing Good quality and Uncoded data, respectively (Table 3).

HBRC also demonstrates a significant level of QC code implementation, employing its own code schema for a substantial portion of its lakes data. Their only parameter that does not undergo QC coding is cyanobacteria data (Figure 7). Similarly, AC, ES and WRC are actively applying QC codes across their lakes parameters. Among these three councils, QC codes are primarily concentrated on ammoniacal nitrogen (NH4N), *Escherichia coli* (ECOLI) and total nitrogen (TN) parameters. It is worth noting that AC and ES use NEMS codes for QC coding purposes, while WRC employs internal codes.

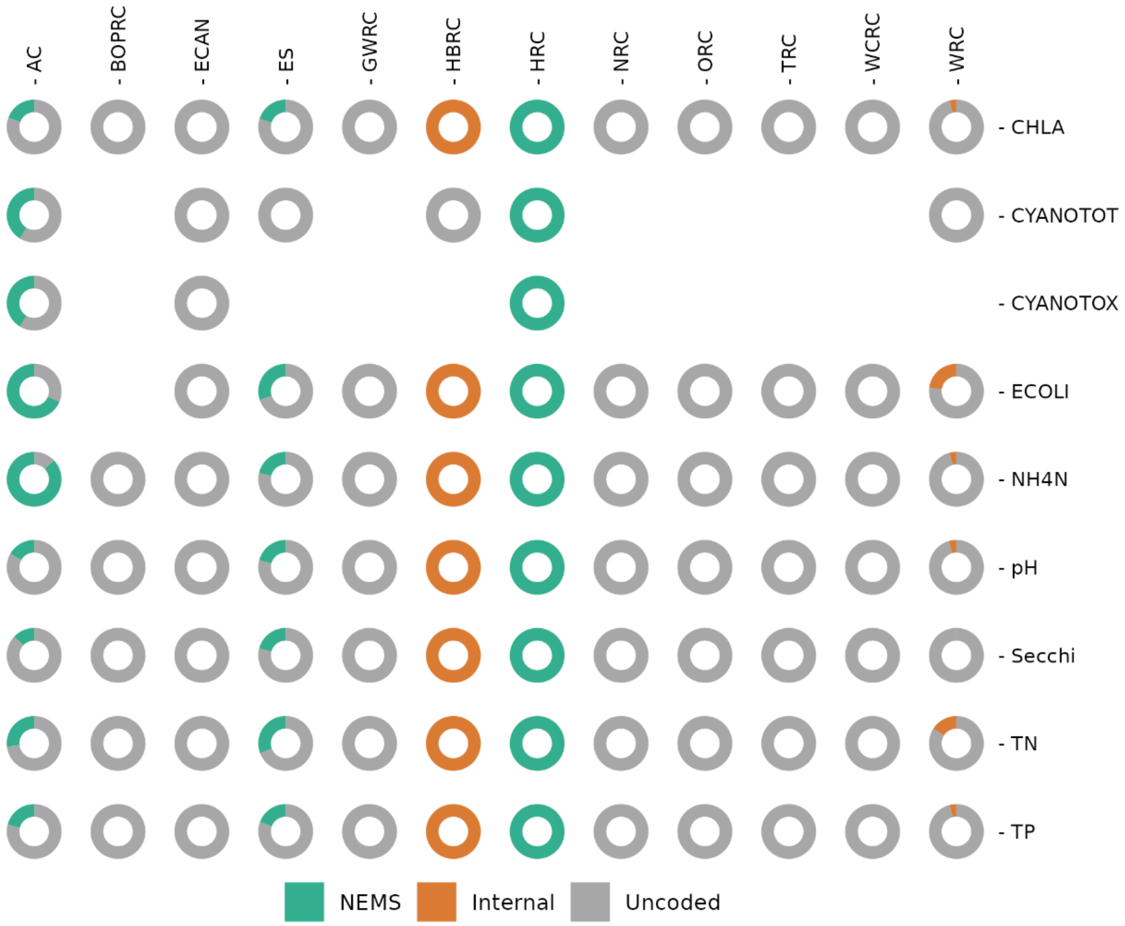


Figure 7. Proportion of records using NEMS QC codes (shown in green) and internal QC codes (shown in orange) for each parameter of the lakes module per council. The proportion of Uncoded records is represented in grey. Parameters that are not measured by councils are presented as missing cells. Councils that do not monitor any lakes in their region are not shown in this figure. For an explanation of abbreviations and terms, see Glossary.

The findings in this section underscore the differing approaches taken by councils in the QC coding process for water quality data analysed in this work. In summary, 11 councils are adopting QC codes to record the quality of their data. Seven of them use codes exclusive to the NEMS schema.

While HRC leads in implementing comprehensive QC coding measures, other councils also display notable efforts in applying QC codes, although with variations in the specific parameters and code schemas used.

In a recent NEMS report (NEMS 2023), a map displays information on the degree of adoption of NEMS standards generally by regional councils and unitary authorities. This map was based on self-reported information provided by councils through a survey, and it indicated that all councils have some level of NEMS implementation.

According to the map, BOPRC appeared to have the highest level of acceptance and full implementation of NEMS standards generally. However, these findings differ from the above results specifically related to adoption of NEMS QC systems.

Hence, while the NEMS implementation report map suggests widespread awareness and implementation of the wider NEMS programme by councils, our analysis focused on adoption and use of NEMS QC codes and metadata and indicates a patchier adoption around the country than is indicated in the broader survey. The data we accessed and analysed were generated during the LAWA 2022 process and did not include any more recent adoption of QC systems. It is also possible that the specific data exchange methods used during the 2022 LAWA update process did not provide access to QC code information for some councils.

4. INVESTIGATING CHANGES IN THE USE OF QC CODE SCHEMAS OVER TIME

In this section, our objective was to examine whether a shift occurred in the percentage of data associated with QC codes following the publication of the initial version of the NEMS quality code schema in 2013. However, our analysis did not reveal any substantial evidence of a significant increase in the adoption of QC codes after this time frame (Figure 8). It is important to acknowledge that our findings do not differentiate between the usage of NEMS and internal QC code systems.

It is noteworthy that some councils already had a significant portion or even all of their data linked with QC codes prior to the publication of the NEMS schema. This suggests that these councils were either aware of the benefits of QC coding and had already implemented their own coding system, or they retrospectively applied QC codes to data collected prior to the introduction of the NEMS QC code schema.

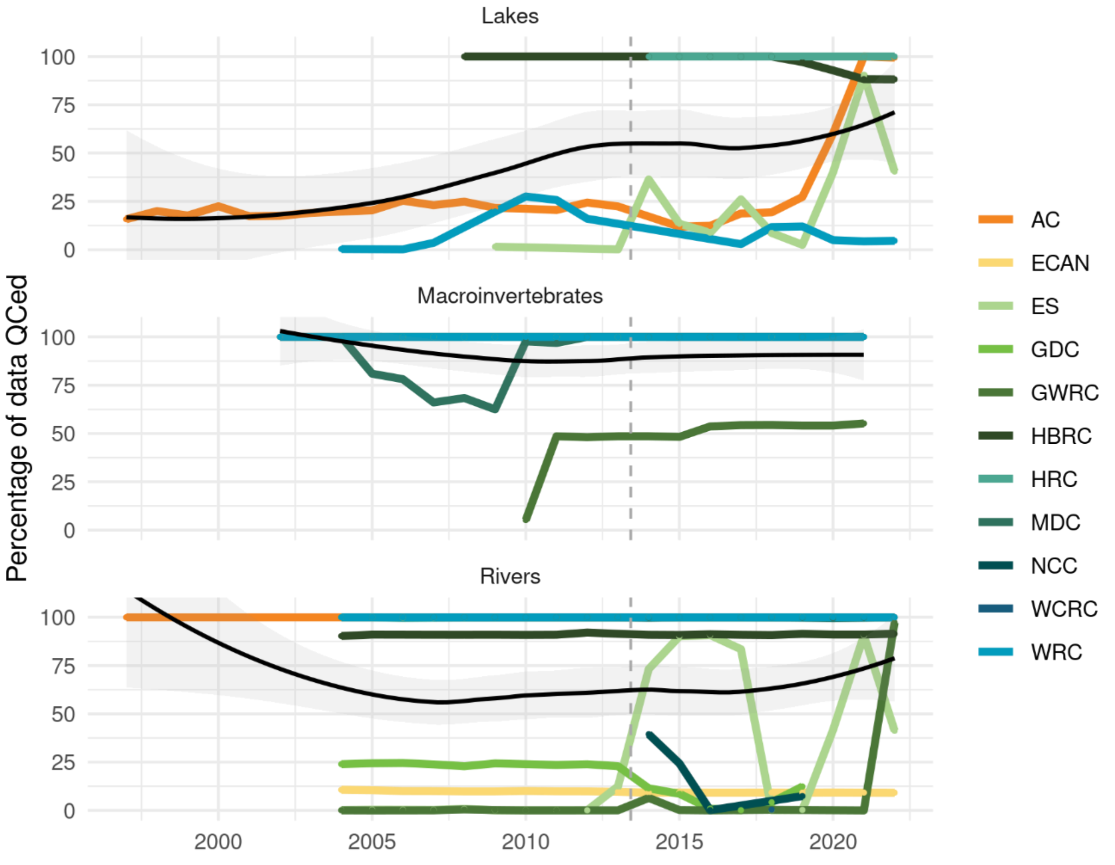
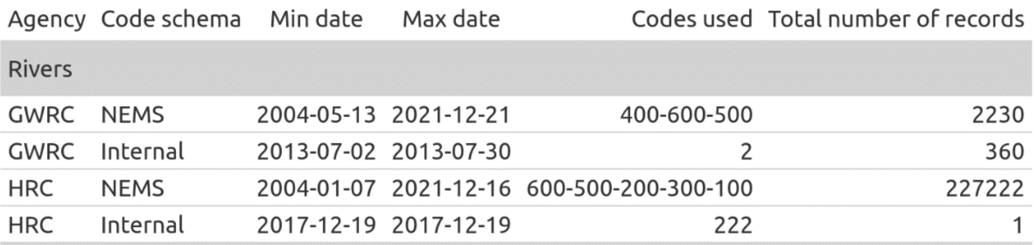


Figure 8. Percentage of records with QC codes per year. The solid black line is a fitted trend line generated using a generalised additive model with a loess smoother. The dashed grey vertical line marks the year when the NEMS national quality code schema report was first published (2013). For an explanation of abbreviations, see Glossary.

We explored the potential to evaluate temporal patterns or shifts between internal and NEMS codes specifically in the case of HRC and GWRC, the two councils that employed a combination of both internal and NEMS code systems. However, after examining the available data, it became evident that these were insufficient to continue with this analysis (Table 4).

HRC had only one record associated with an internal code (222), and without further information we refrained from making any assumptions regarding its relationship to the NEMS code 200. Similarly, GWRC had only one month of data using internal codes (Table 4).

Table 4. Councils using a combination of code schemas to QC their data and the duration for which each of these schemas was in use.



Agency	Code schema	Min date	Max date	Codes used	Total number of records
Rivers					
GWRC	NEMS	2004-05-13	2021-12-21	400-600-500	2230
GWRC	Internal	2013-07-02	2013-07-30	2	360
HRC	NEMS	2004-01-07	2021-12-16	600-500-200-300-100	227222
HRC	Internal	2017-12-19	2017-12-19	222	1

5. EXPLORING DIFFERENCES IN WATER QUALITY DATA ASSOCIATED WITH DIFFERENT QC CODES

To investigate the differences between data with different QC codes, we classified the data into four groups: Uncoded*,[5] Synthetic, Good quality and Poor quality (Table 5). As the internal codes could not be translated directly into Poor quality and Good quality data, the records associated with internal codes were removed from this analysis. It is worth noting that not all modules had records represented in all four QC code groups.

Table 5. Mapping table for classifying QC codes into four qualitative groups. Note that for the analyses in this section we combined Uncoded data with NEMS 200 into a single QC group called Uncoded*.

QC code group	Codes
Uncoded*	200 and Uncoded
Good quality	600, 500, 610, 550, 543, 560
Synthetic	300
Poor quality	400, 460, 450, 100, 403, 404, 405

We used boxplots to visualise the distribution of data values across the QC code groups for rivers, macroinvertebrates and lakes data. Although we did not observe any major differences between the QC code groups (Figure 9, Figure 10, Figure 11), we did notice that the data distribution varied considerably depending on the parameter being analysed and so conducted further investigation as outlined below.

[5] For the analyses in this section, we combined Uncoded data with NEMS 200 into a single QC group called Uncoded*.

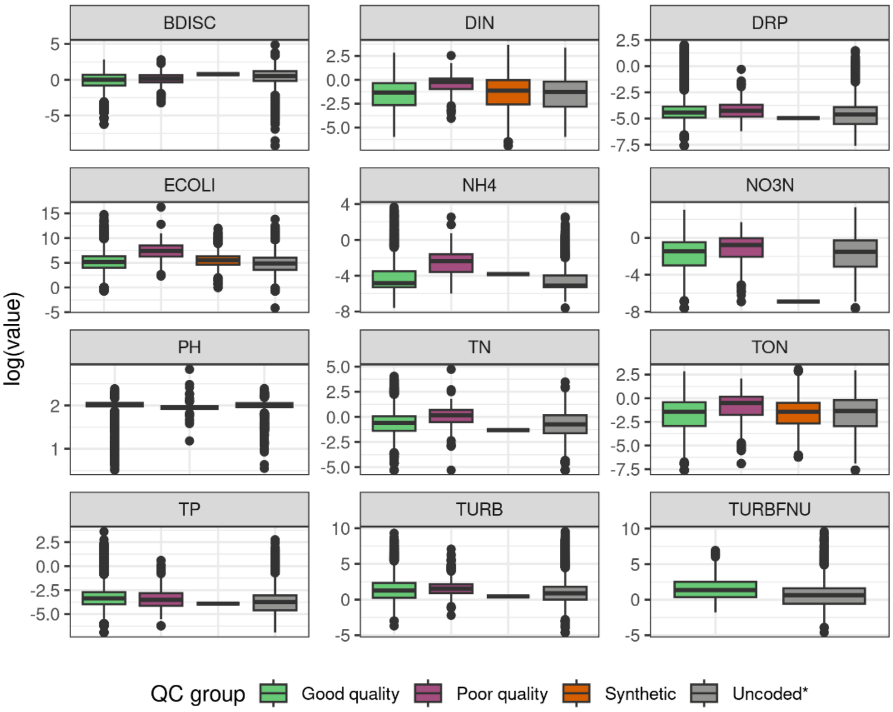


Figure 9. Distribution of values from all rivers water quality parameters. Note that the values were log-transformed to improve the visualisation of the data. For an explanation of abbreviations and terms, see Glossary.

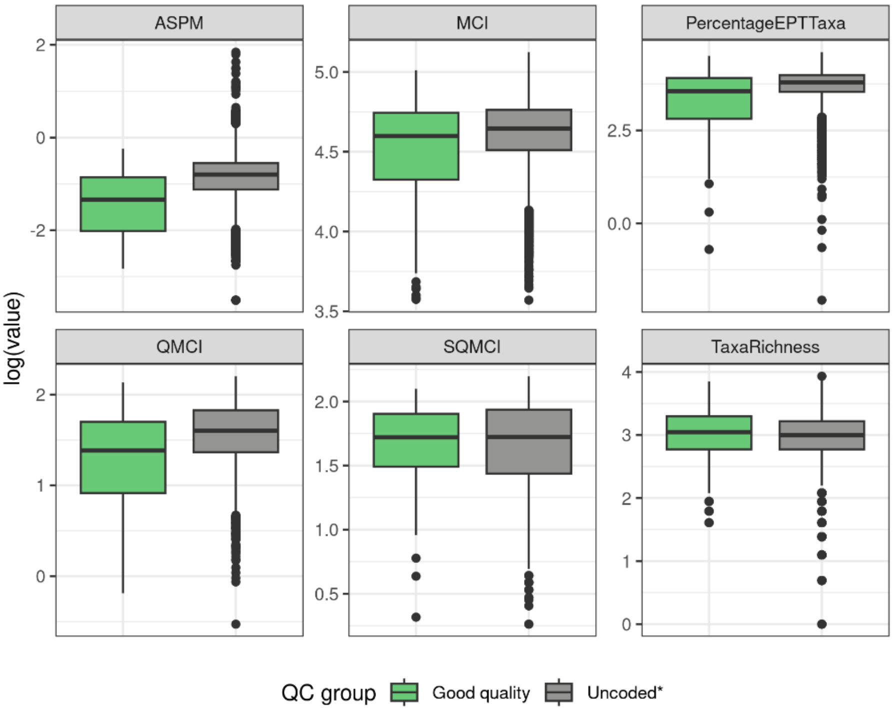


Figure 10. Distribution of values from all macroinvertebrate parameters. Note that the values were log-transformed to improve the visualisation of the data. For an explanation of abbreviations and terms, see Glossary.

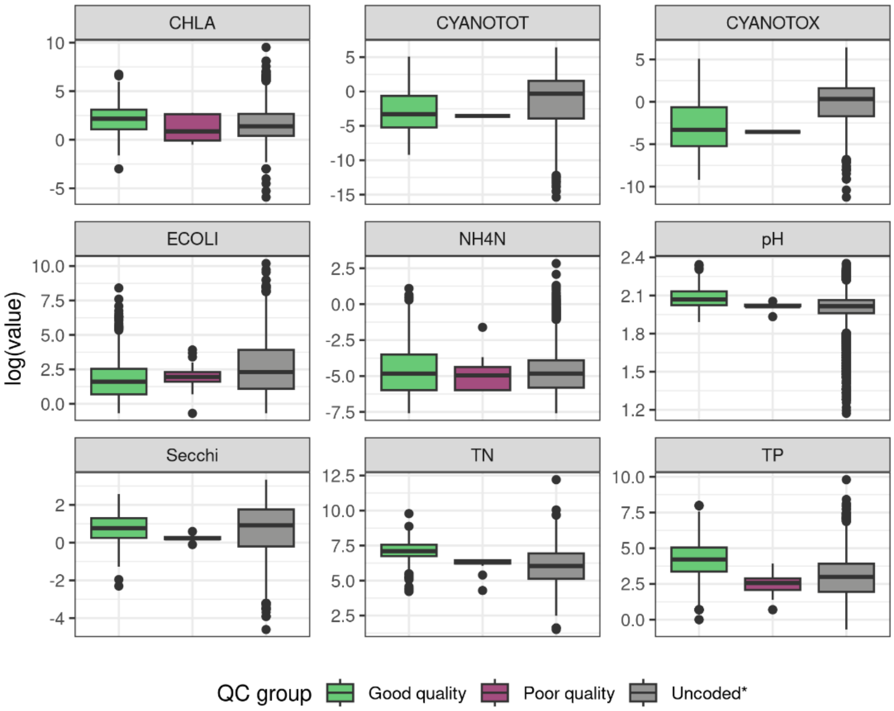


Figure 11. Distribution of values from all lakes water quality parameters. Note that the values were log-transformed to improve the visualisation of the data. For an explanation of abbreviations and terms, see Glossary.

It is worth noting that the number of records in each QC group was highly variable, with very few records associated with Poor quality codes (Table 6, Table 7, Table 8). This limitation should be considered when interpreting the results from this section.

Table 6. Number of data records for rivers water quality parameters for each QC code group. For an explanation of abbreviations and terms, see Glossary.

Parameter	Poor quality	Good quality	Synthetic	Uncoded*
BDISC	1,086	23,661	2	58,335
DIN	113	11,558	37,757	50,368
DRP	182	36,045	1	72,814
ECOLI	110	33,707	2,543	76,298
NH4	137	35,975	1	74,358
NO3N	354	27,382	1	59,385
PH	174	28,324	NA	76,845
TN	145	28,693	1	67,080
TON	466	32,569	3,086	67,142
TP	547	31,369	1	69,349
TURB	917	34,637	1	70,664
TURBFNU	NA	1,139	NA	4,894

Table 7. Number of data records for macroinvertebrate parameters for each QC code group. For this module, no records were associated with the Synthetic and Poor quality QC groups. For an explanation of abbreviations and terms, see Glossary.

Parameter	Good quality	Uncoded*
ASPM	750	10,817
MCI	1,117	13,283
PercentageEPTTaxa	1,114	11,804
QMCI	750	9,335
SQMCI	337	2,817
TaxaRichness	1,115	12,874

Table 8. Number of data records for lakes water quality parameters for each QC code group. For this module, no records were associated with the Synthetic QC group. For an explanation of abbreviations and terms, see Glossary.

Parameter	Poor quality	Good quality	Uncoded*
CHLA	18	595	14,778
CYANOTOT	1	280	2,494
CYANOTOX	1	280	1,079
ECOLI	18	709	4,991
NH4N	11	951	12,010
Secchi	13	246	9,956
TN	12	558	13,945
TP	49	580	14,733
pH	9	335	10,457

To further explore whether there were substantial differences in the distribution of data among the QC code groups, we focused on *Escherichia coli* (ECOLI) for rivers, Macroinvertebrate Community Index (MCI) for macroinvertebrates and total phosphorus (TP) for lakes, and compared the data distribution among groups using regression models. This analysis serves to compare and test whether there is a significant difference in the values associated with each pair of QC code categories. Understanding these distinctions among the data distribution in different QC code categories can underline the importance of accounting for such differences in subsequent analysis.

For all parameters we fitted a generalised linear regression model with a gamma distribution as the error distribution for the models. A gamma distribution is suitable for positive and continuous data and can address heteroskedasticity in the data. After fitting the model for each parameter, we conducted an autocorrelation test on the model residuals. The results of the test showed evidence of temporal correlation in the residuals, indicating that neighbouring observations were not independent. This suggests that there may be underlying temporal patterns or trends in the data that were not captured by the models. Therefore, it is important to consider this temporal correlation when interpreting the results, as well as the variable number of records in each QC code class, mentioned previously, as these may affect the statistical power of the models.

The pairwise comparisons between the QC code groups applied to the models revealed that, for TP values, there was no significant difference in the data distribution between Good quality data and Poor quality data based on the *p*-value of the post-hoc test (*p*-value = 0.804). However, the distribution of the data was significantly different between Uncoded* and Good quality data and between Poor quality and Uncoded* data (*p*-value < 0.001, for both cases).

On other hand, the results for the ECOLI model indicated significant differences between values classified as Good quality and Poor quality (*p*-value = 0.020) and between Poor quality and Uncoded* ECOLI values (*p*-value = 0.043). However, no significant differences were observed between Good quality and Uncoded* values (*p*-value = 0.355) and Poor quality values and Synthetic values (*p*-value = 0.628).

Macroinvertebrate data had records represented only in the Good quality and Uncoded QC code groups. However, the statistical model applied to this module did not indicate a significant difference between the MCI values associated with these two groups (*p*-value = 0.114).

The results presented in this section reveal variations in the distribution of data across different QC groups for certain parameters, which would have an impact on subsequent analyses. For example, NPS-FM band assignments may yield differing results when considering the entire dataset as opposed to using only data classified as Good quality through QC codes. Conversely, for some parameters there was no evidence for differences in the data distribution between data classified as Good quality and Poor quality.

When considering these results, it is worthwhile considering potential reasons why data may be classified as Poor quality data. Reasons could include a faulty field meter, accidental sample contamination, or delayed transit to the laboratory, meaning the results are unreliable and unlikely to reflect the real situation. Some data are classified as Poor quality because they were collected outside the NEMS sampling guideline period (e.g. insufficient wait period since the last flood). In these situations, the data are reliable but should not be compared directly with data collected according to guidelines. It is also possible that some data are classified as Poor quality due to several relatively minor variations from NEMS requirements (e.g. lack of recent training of staff, absence of information on exact time of sampling). In these situations, the data may be reliable but should be treated with caution.

Decisions on whether to exclude Poor quality or Uncoded data from subsequent analyses should be made cautiously. It is not appropriate to include data that are unreliable, but on the other hand valuable information will be lost if data downgraded by relatively minor issues, or collected to a high standard but before QC systems were in place, are excluded from analyses.

One potential solution to address this issue could involve the use of child QC codes to specify which data councils recommend should be excluded from future analyses due to significant data collection errors. It is essential to document the reasons why specific data receive a particular code within the metadata. Moreover, NEMS could play a role in advising councils on how to flag such cases within their QC codes.

6. METADATA AVAILABILITY AND COMPOSITION

In this section we present only metadata relating to river water quality, as the metadata patterns across all modules were found to be quite similar, leading to similar conclusions. However, for comprehensive information, all figures and tables related to lakes and macroinvertebrates can be found in Appendices 3 to 12.

Figure 12 provides an overview of the proportion of records that have associated metadata information per council and demonstrates that most councils were recording some type of metadata with their data. All councils, except TDC, have some form of metadata associated with their river water quality data. Seven of the councils we successfully extracted metadata from have metadata attached to all their data records. Three other councils have metadata attached to more than 95% of their data, while two councils (BOPRC and WCRC) had metadata attached to less than 5% of their records.

When comparing the outcomes of the other two modules (macroinvertebrates – Appendix 4, and lakes – Appendix 9), it becomes apparent that seven councils (ECAN, ES, GWRC, HBRC, HRC, NRC and TRC) have implemented metadata recording to some extent for data from all modules.

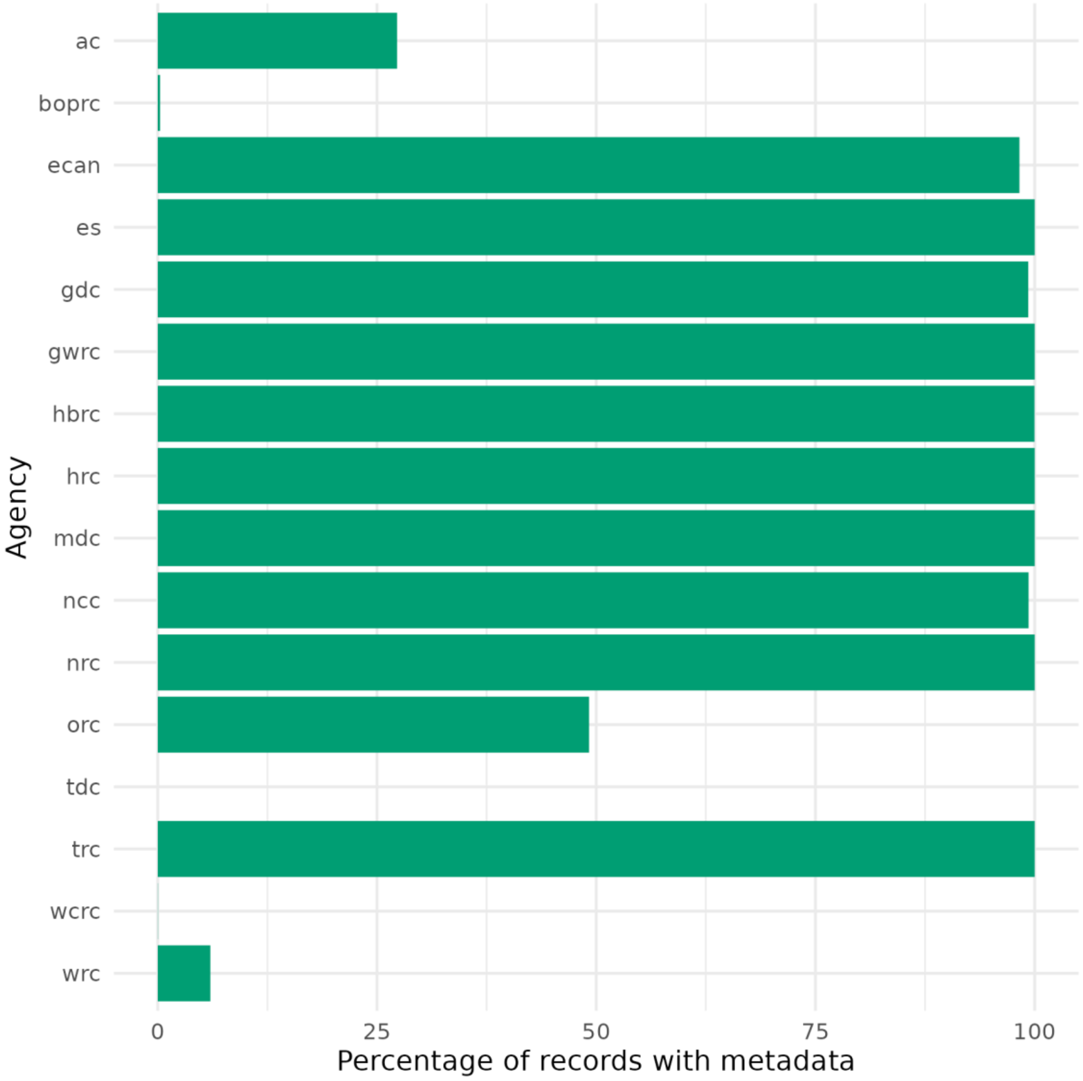
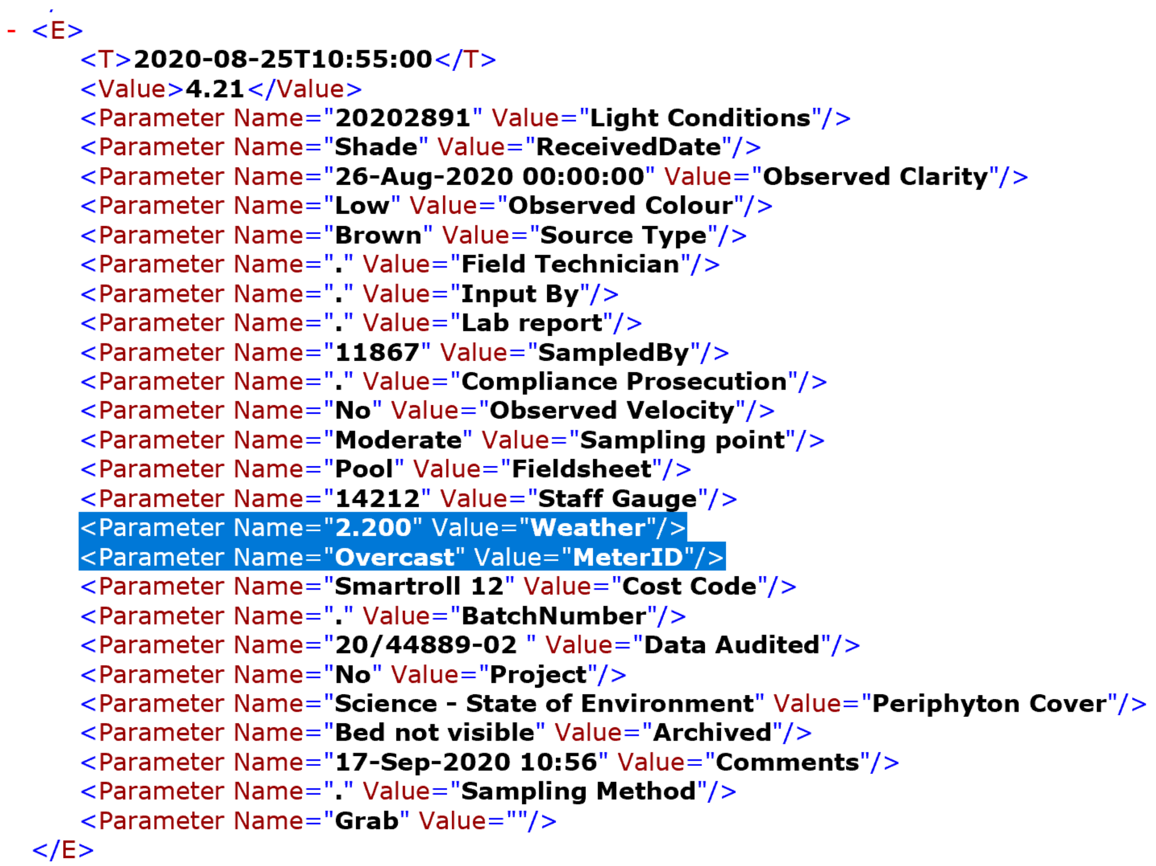


Figure 12. Proportion of river water quality data with associated metadata per council. Only councils that provide their data through a data server were included in this analysis. All councils, except TDC, have some form of metadata associated with their river water quality data. It is worth noting that less than 5% of BOPRC and WCRC records include metadata. We bring attention to this as the histogram figure might erroneously give the impression that they do not have any data with associated metadata. For an explanation of abbreviations, see Glossary.

We identified a substantial number of metadata variables for rivers (195, excluding HRC), macroinvertebrates (86) and lakes (178). We did not include HRC's unique metadata variables in this descriptive summary as there was a mismatch between how the name of the metadata variable and the actual value are stored (Figure 13).



```
- <E>
      <T>2020-08-25T10:55:00</T>
      <Value>4.21</Value>
      <Parameter Name="20202891" Value="Light Conditions"/>
      <Parameter Name="Shade" Value="ReceivedDate"/>
      <Parameter Name="26-Aug-2020 00:00:00" Value="Observed Clarity"/>
      <Parameter Name="Low" Value="Observed Colour"/>
      <Parameter Name="Brown" Value="Source Type"/>
      <Parameter Name="." Value="Field Technician"/>
      <Parameter Name="." Value="Input By"/>
      <Parameter Name="." Value="Lab report"/>
      <Parameter Name="11867" Value="SampledBy"/>
      <Parameter Name="." Value="Compliance Prosecution"/>
      <Parameter Name="No" Value="Observed Velocity"/>
      <Parameter Name="Moderate" Value="Sampling point"/>
      <Parameter Name="Pool" Value="Fieldsheet"/>
      <Parameter Name="14212" Value="Staff Gauge"/>
      <Parameter Name="2.200" Value="Weather"/>
      <Parameter Name="Overcast" Value="MeterID"/>
      <Parameter Name="Smartroll 12" Value="Cost Code"/>
      <Parameter Name="." Value="BatchNumber"/>
      <Parameter Name="20/44889-02 " Value="Data Audited"/>
      <Parameter Name="No" Value="Project"/>
      <Parameter Name="Science - State of Environment" Value="Periphyton Cover"/>
      <Parameter Name="Bed not visible" Value="Archived"/>
      <Parameter Name="17-Sep-2020 10:56" Value="Comments"/>
      <Parameter Name="." Value="Sampling Method"/>
      <Parameter Name="Grab" Value=""/>
  </E>
```

Figure 13. Example of HRC metadata (from the original XML data retrieved from the HRC data server), where the variable name of the metadata (which should be informed in the Parameter Name attribute) is switched with the actual value of the previous metadata record.

The large number of metadata variables used by councils (shown as a word cloud in Figure 14) indicates a significant lack of consistency in how this information is captured and managed, both within individual councils and across different councils. This inconsistency highlights the need for standardised approaches to ensure greater uniformity and interoperability in metadata practices among councils.

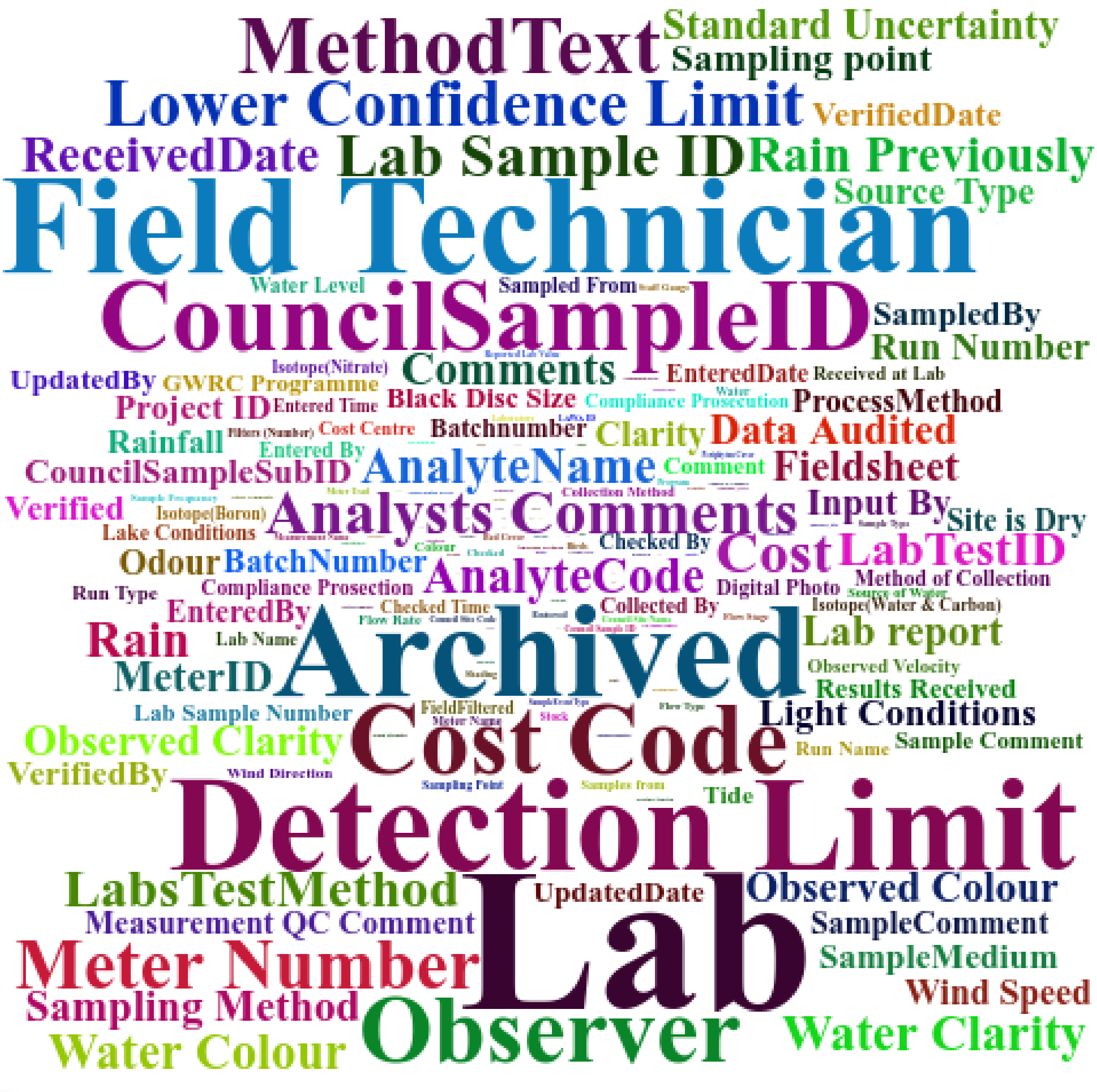


Figure 14. Word cloud with all the different metadata variables associated to the river water quality data. The word size is related to the frequency at which each metadata variable was used.

To examine possible differences in the type of metadata used, we analysed the 10 variables that appeared most frequently for each council and for at least 10% of the records. As shown in Figure 15, the metadata for all parameters typically falls into one of three categories: specific sample information (such as project name, technician and sample ID), weather conditions (such as rainfall), and analyses (methods, laboratory, detection limit, black disc size). The type of metadata also varies with the water quality parameter (Figure 16), where lab and method appear as the most common metadata for dissolved reactive phosphorus (DRP), *Escherichia coli* (ECOLI), ammoniacal nitrogen (NH4N), nitrate nitrogen (NO3N), total nitrogen (TN), total oxidised nitrogen

(TON), total phosphorus (TP) and turbidity (TURB). It is worth noting that the only metadata fields available in non-Hilltop servers were the QC codes and information related to censored values. This might indicate that additional metadata from these councils are stored in another data source and could not be accessed in this study.

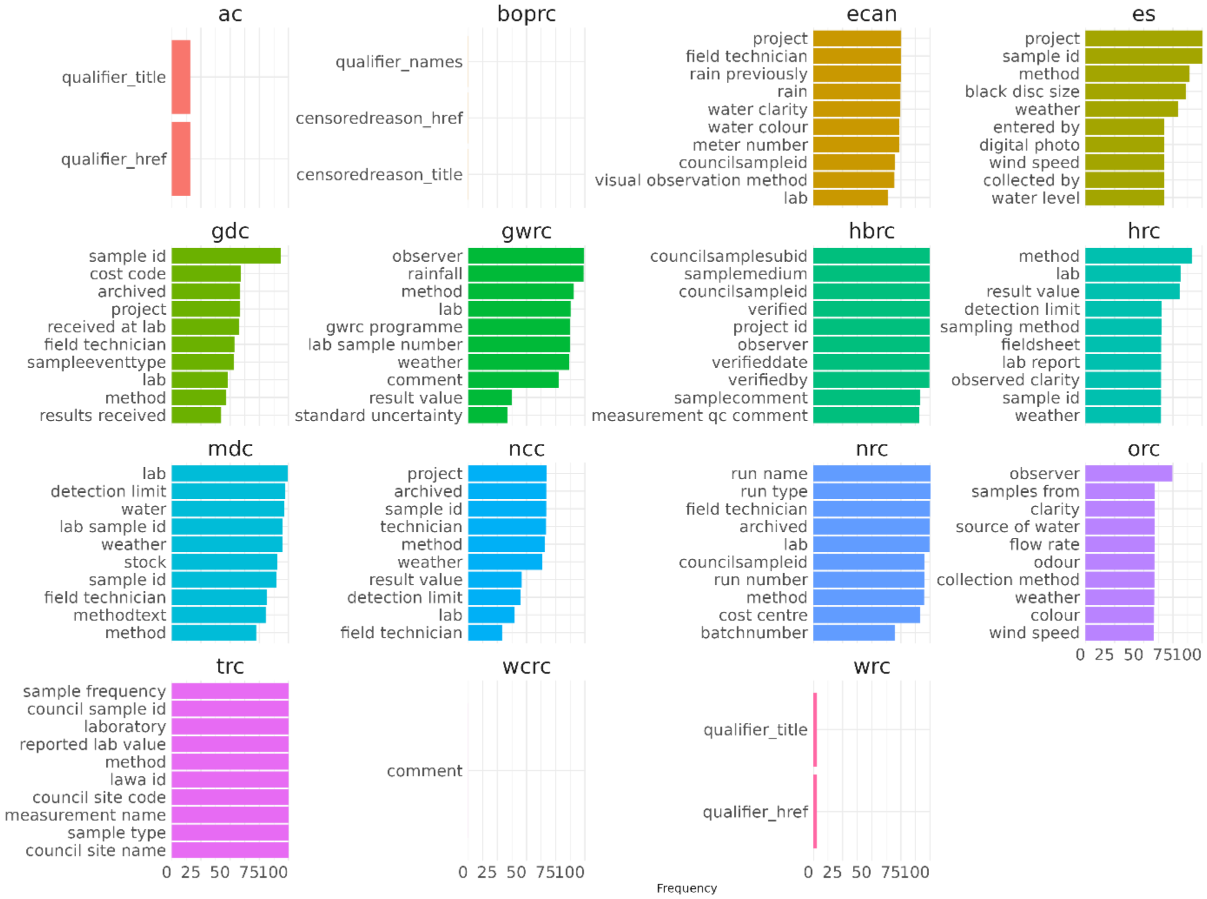


Figure 15. Proportion of records associated with the 10 most frequent metadata variables used by each council in the river water quality data. For an explanation of abbreviations, see Glossary.

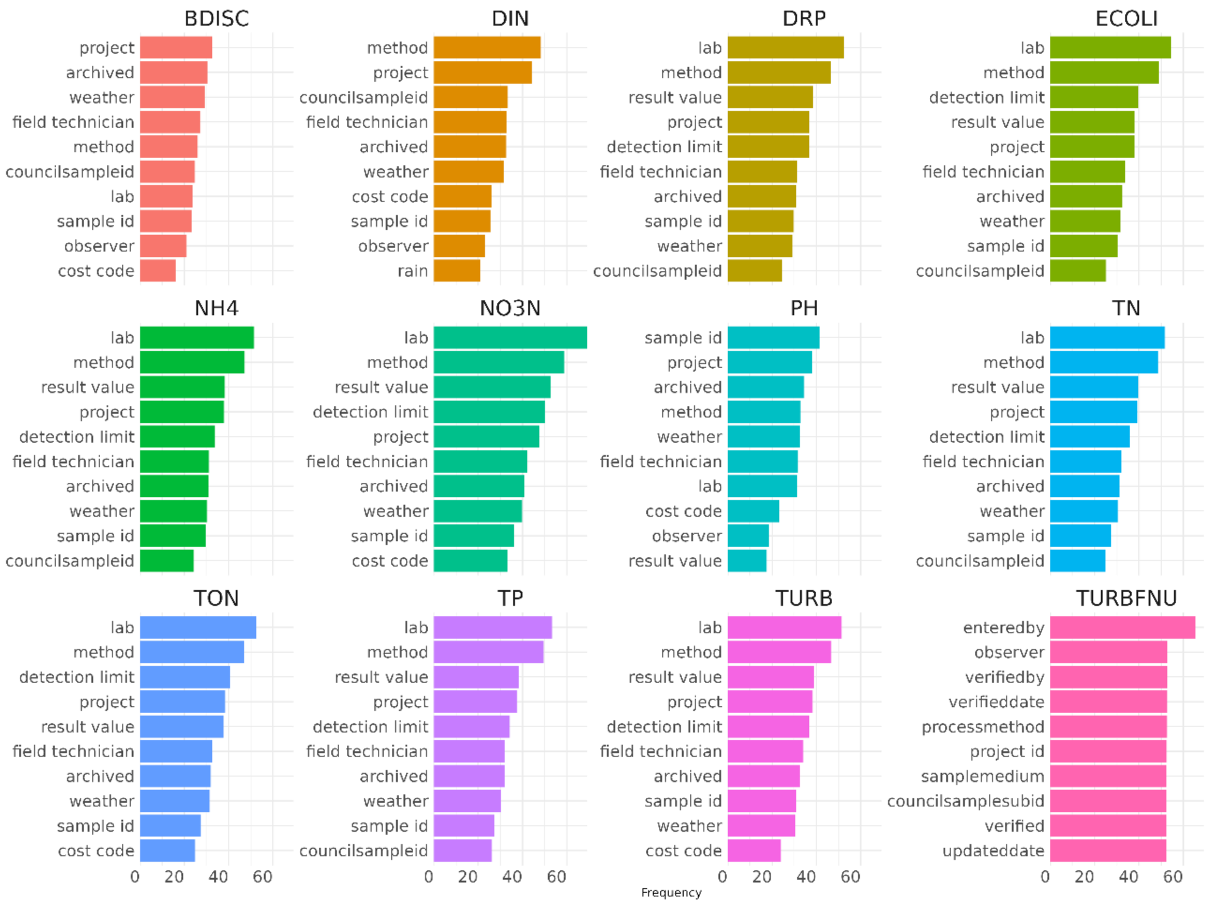


Figure 16. Proportion of records associated to the 10 most frequent metadata variables used in each river water quality variable. For an explanation of abbreviations and terms, see Glossary.

Laboratory methods play a crucial role in ensuring the quality of subsequent data analysis. Recognising their significance, we specifically focused on metadata variables associated with laboratory analyses for this last analysis of the metadata. We observed that ‘detection limit’, ‘lab’ and ‘method’ were the most frequently used metadata variables relevant to the laboratory analysis of the data. The ‘detection limit’ metadata provides information about the minimum detectable concentration for a given parameter, while the ‘lab’ variable stores details about the specific laboratory that conducted the analysis. In contrast, the ‘method’ variable contains more comprehensive information about the specific procedures and techniques employed during the laboratory analysis.

However, it is important to note that there is considerable variability in how this information is recorded (Figure 17). For instance, terms such as ‘hill’, ‘hills’, ‘hill laboratories’ and ‘chch’ may be used interchangeably to represent the laboratory in the ‘lab’ metadata variable. This inconsistency in metadata recording presents challenges in collating and harmonising the recorded data effectively.

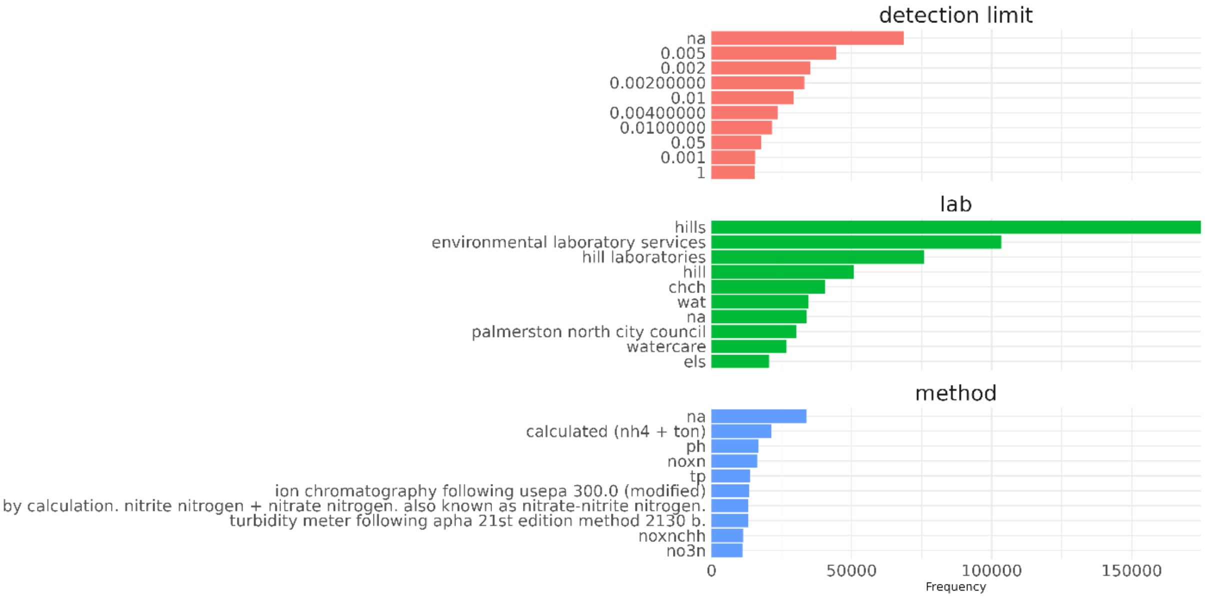


Figure 17. Number of records associated to the most frequent metadata values for three metadata variables relevant to the laboratory analysis of the river water quality data.

7. SUMMARY AND RECOMMENDATIONS

The first national attempt to collate QC code and metadata information stored in councils' public data servers is described in this report. This work provides a picture of how councils are currently using QC codes and storing relevant metadata alongside their water quality data and provides an understanding of levels of QC code and metadata implementation. We hope this will lead to discussion about opportunities for improving data consistency and subsequent analyses.

In Section 3 we presented general summaries of how QC codes and QC schemas are being used in different councils, modules and individual parameters. Based on the water quality data that was sourced, 11 of 16 councils surveyed have adopted QC codes to some extent. However, only half of the river water quality data sourced had been assigned a QC code, and implementation of QC coding was even less common for macroinvertebrate data (25%) and lake health data (13%). For the data that do have QC codes, about half use internal council QC codes rather than the NEMS coding system.

In Section 4 we investigated the adoption of QC codes throughout the years for each council. Results showed that the percentage of data with QC codes varied over the years across councils with no clear pattern or indication that the adoption of QC codes is increasing with time. Our results also indicated that QC codes can be applied retrospectively, as some councils already had a significant portion, or even all, of their data linked with QC codes prior to the publication of the NEMS schema.

In Section 5 we examined whether data with different QC codes exhibit differences in descriptive statistics. For *Escherichia coli* (ECOLI) measured in rivers, we found that the data distribution of values classified as Good quality and Poor quality and between Poor quality and Uncoded* are significantly different. Conversely, for total phosphorus (TP) in lakes and Macroinvertebrate Community Index (MCI) we found that the data distribution between records classified as Good quality and Poor quality data and Poor quality and Uncoded* are not significantly different. Decisions based on QC codes on whether to exclude Poor quality or Uncoded data from subsequent analyses should be made cautiously. It is not appropriate to include data that are unreliable, but on the other hand valuable information will be lost if data downgraded by relatively minor issues, or collected to a high standard but before QC systems were in place, are excluded from analyses. The use of child QC codes to specify which data councils recommend should be excluded from future analyses may be a solution to this issue.

In Section 6 we presented a general overview of the metadata stored in the dataset analysed in this project. The results in this section showed that fifteen councils have recorded metadata alongside their waterway health data, and seven of these had some level of metadata associated with records across all modules. These metadata cover a very wide range of topics, including administrative information (e.g. project name, sample ID, field technician name), weather conditions at the time of sampling, field instrument details and laboratory sample analysis methods (e.g. laboratory name, analysis method, detection limit). We identified a substantial number of metadata variables for rivers (> 195), macroinvertebrates (86) and lakes (178), indicating a significant lack of consistency in how councils capture and manage this information. This inconsistency highlights the need for standardised approaches to ensure greater uniformity in metadata practices among councils.

Our analyses provide a snapshot view of the adoption of QC codes and use of metadata, and highlight the current high level of variability in adoption among councils and across different datasets. At some councils, the implementation of QC codes appears to have varied over time, with bursts of activity followed by quieter periods. This contrasts with a slow and steady adoption of the NEMS QC code system over time, as might have been expected.

In general, the level of adoption of QC codes is relatively low. Consideration should be given to how adoption can be enhanced. This report may be a catalyst for further adoption among councils that are just starting on this journey. There may be value in distilling lessons from those councils that are well advanced in their adoption and use of QC codes and metadata, to assist other councils with their efforts. Other mechanisms requiring the use of QC codes may also need to be considered given that voluntary adoption appears to have had mixed success.

Below we list key recommendations identified from the results presented in this report.

1. The high level of variability in adoption of QC coding to waterway health data among councils and across different datasets indicates that there have been some barriers to adoption for some councils. Further work in understanding the factors influencing QC code adoption across all councils is needed to identify these barriers and help overcome them.

2. NEMS (2016) recommends that individual data records should be stored with information on ‘units’. However, the metadata that we extracted rarely presented units alongside the individual data records; instead, the ‘units’ information is usually available in the overall header of the time-series response. As this is critical information when analysing data, we recommend that councils present the units as part of their metadata per record.

3. As recommended by NEMS (2016), metadata that was extracted included information on laboratory name, sampling location and details on test methods. However, there needs to be more consistency in how these metadata variables are labelled in councils’ servers and how metadata values are recorded. Such standardisation will

improve how water quality data is managed across Aotearoa New Zealand, assist data analysis and enable more impactful research using these data.

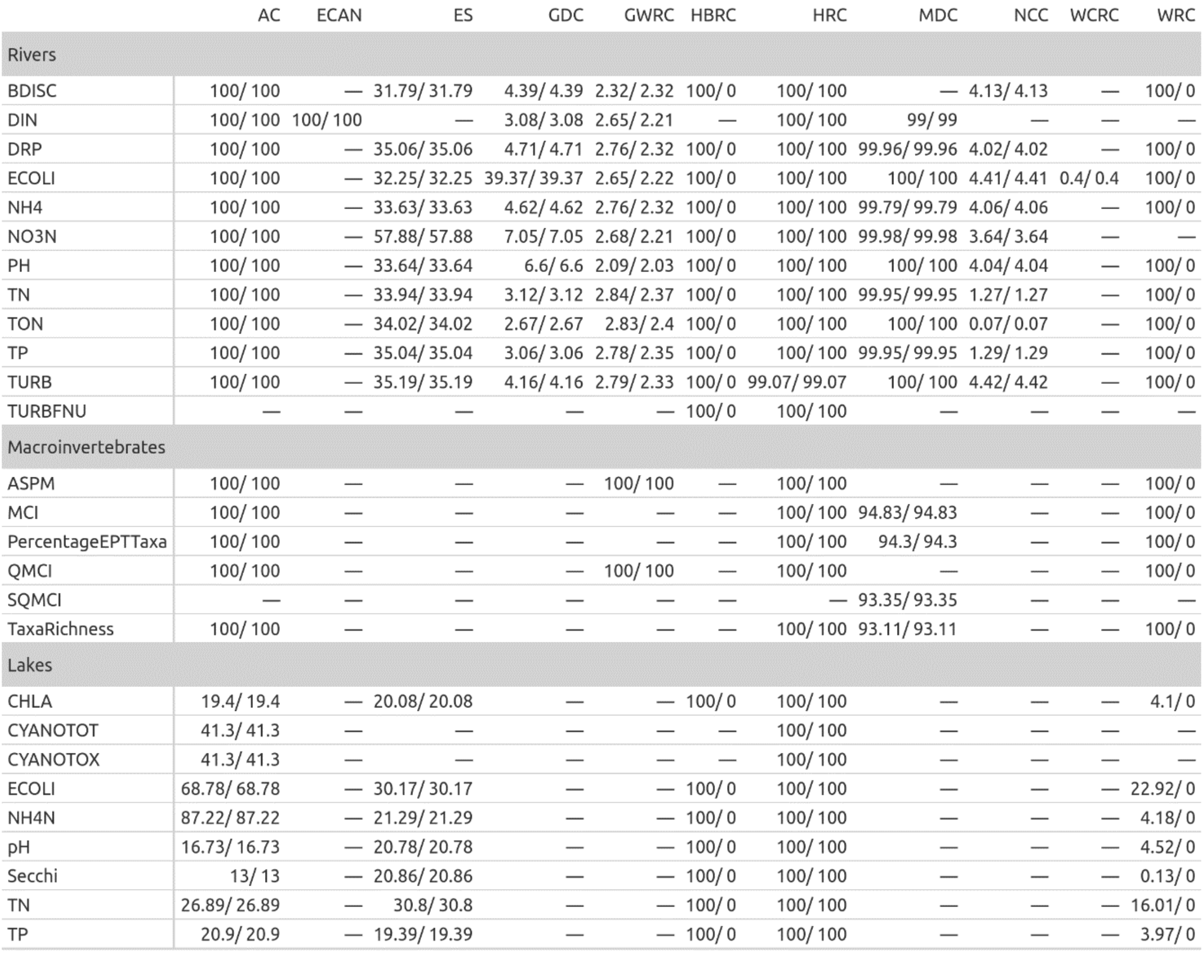
4. Councils are constantly updating the way they publish their data and making improvements to their data servers. Data migration and changes to servers can make it challenging for end-users to interact with the data on a repeat basis. We understand that there is an initiative proposed to centralise the storage of environmental data for all councils, rather than relying on the federated set of data servers that currently store this information. We support the move to centralise data storage, although individual councils will still need to be ultimately responsible for the quality and completeness of the data that are provided.

5. Data from all councils using data servers are retrieved in XML format. This is a structured and widely used data format, however, due to its hierarchical and verbose characteristics it is often perceived as more complex when compared to more contemporary and user-friendly data structures, such as JavaScript Object Notation (JSON). We recommend the transition to JSON responses as the centralised national data server is developed and implemented in the future. Although not critical, this change would benefit end-users who access council servers directly by an application programming interface endpoint connection to request water quality data, enhancing data accessibility and alignment with current data standards.

6. Decisions on whether to exclude Poor quality or Uncoded data from subsequent analyses should be made cautiously and based upon knowledge of why the data were given these QC codes. We recommend that guidance is developed on the use of child QC codes to specify which data should be excluded from future analyses.

8. APPENDICES

Appendix 1. Proportion of records using QC codes per parameter and per council. The first number refers to total percentage of records with QC codes from any QC schema, and the number after '/' indicates the percentage of records using codes from the NEMS schema. Note that only councils that use QC codes are represented in this figure. For an explanation of abbreviations and terms, see Glossary.



	AC	ECAN	ES	GDC	GWRC	HBRC	HRC	MDC	NCC	WCRC	WRC
Rivers											
BDISC	100/ 100	—	31.79/ 31.79	4.39/ 4.39	2.32/ 2.32	100/ 0	100/ 100	—	4.13/ 4.13	—	100/ 0
DIN	100/ 100	100/ 100	—	3.08/ 3.08	2.65/ 2.21	—	100/ 100	99/ 99	—	—	—
DRP	100/ 100	—	35.06/ 35.06	4.71/ 4.71	2.76/ 2.32	100/ 0	100/ 100	99.96/ 99.96	4.02/ 4.02	—	100/ 0
ECOLI	100/ 100	—	32.25/ 32.25	39.37/ 39.37	2.65/ 2.22	100/ 0	100/ 100	100/ 100	4.41/ 4.41	0.4/ 0.4	100/ 0
NH4	100/ 100	—	33.63/ 33.63	4.62/ 4.62	2.76/ 2.32	100/ 0	100/ 100	99.79/ 99.79	4.06/ 4.06	—	100/ 0
NO3N	100/ 100	—	57.88/ 57.88	7.05/ 7.05	2.68/ 2.21	100/ 0	100/ 100	99.98/ 99.98	3.64/ 3.64	—	—
PH	100/ 100	—	33.64/ 33.64	6.6/ 6.6	2.09/ 2.03	100/ 0	100/ 100	100/ 100	4.04/ 4.04	—	100/ 0
TN	100/ 100	—	33.94/ 33.94	3.12/ 3.12	2.84/ 2.37	100/ 0	100/ 100	99.95/ 99.95	1.27/ 1.27	—	100/ 0
TON	100/ 100	—	34.02/ 34.02	2.67/ 2.67	2.83/ 2.4	100/ 0	100/ 100	100/ 100	0.07/ 0.07	—	100/ 0
TP	100/ 100	—	35.04/ 35.04	3.06/ 3.06	2.78/ 2.35	100/ 0	100/ 100	99.95/ 99.95	1.29/ 1.29	—	100/ 0
TURB	100/ 100	—	35.19/ 35.19	4.16/ 4.16	2.79/ 2.33	100/ 0	99.07/ 99.07	100/ 100	4.42/ 4.42	—	100/ 0
TURBFNU	—	—	—	—	—	100/ 0	100/ 100	—	—	—	—
Macroinvertebrates											
ASPM	100/ 100	—	—	—	100/ 100	—	100/ 100	—	—	—	100/ 0
MCI	100/ 100	—	—	—	—	—	100/ 100	94.83/ 94.83	—	—	100/ 0
PercentageEPTTaxa	100/ 100	—	—	—	—	—	100/ 100	94.3/ 94.3	—	—	100/ 0
QMCI	100/ 100	—	—	—	100/ 100	—	100/ 100	—	—	—	100/ 0
SQMCI	—	—	—	—	—	—	—	93.35/ 93.35	—	—	—
TaxaRichness	100/ 100	—	—	—	—	—	100/ 100	93.11/ 93.11	—	—	100/ 0
Lakes											
CHLA	19.4/ 19.4	—	20.08/ 20.08	—	—	100/ 0	100/ 100	—	—	—	4.1/ 0
CYANOTOT	41.3/ 41.3	—	—	—	—	—	100/ 100	—	—	—	—
CYANOTOX	41.3/ 41.3	—	—	—	—	—	100/ 100	—	—	—	—
ECOLI	68.78/ 68.78	—	30.17/ 30.17	—	—	100/ 0	100/ 100	—	—	—	22.92/ 0
NH4N	87.22/ 87.22	—	21.29/ 21.29	—	—	100/ 0	100/ 100	—	—	—	4.18/ 0
pH	16.73/ 16.73	—	20.78/ 20.78	—	—	100/ 0	100/ 100	—	—	—	4.52/ 0
Secchi	13/ 13	—	20.86/ 20.86	—	—	100/ 0	100/ 100	—	—	—	0.13/ 0
TN	26.89/ 26.89	—	30.8/ 30.8	—	—	100/ 0	100/ 100	—	—	—	16.01/ 0
TP	20.9/ 20.9	—	19.39/ 19.39	—	—	100/ 0	100/ 100	—	—	—	3.97/ 0

Appendix 2. URL addresses used to source data from councils' data servers and the type of data server used by each council. Missing cells in the report indicate that the data for the respective module are not available in the council's data server and therefore could not be included in this analysis. Note that additional *site* and *variable name* must be specified at the end of each URL to return a valid data response. For an explanation of abbreviations, see Glossary.

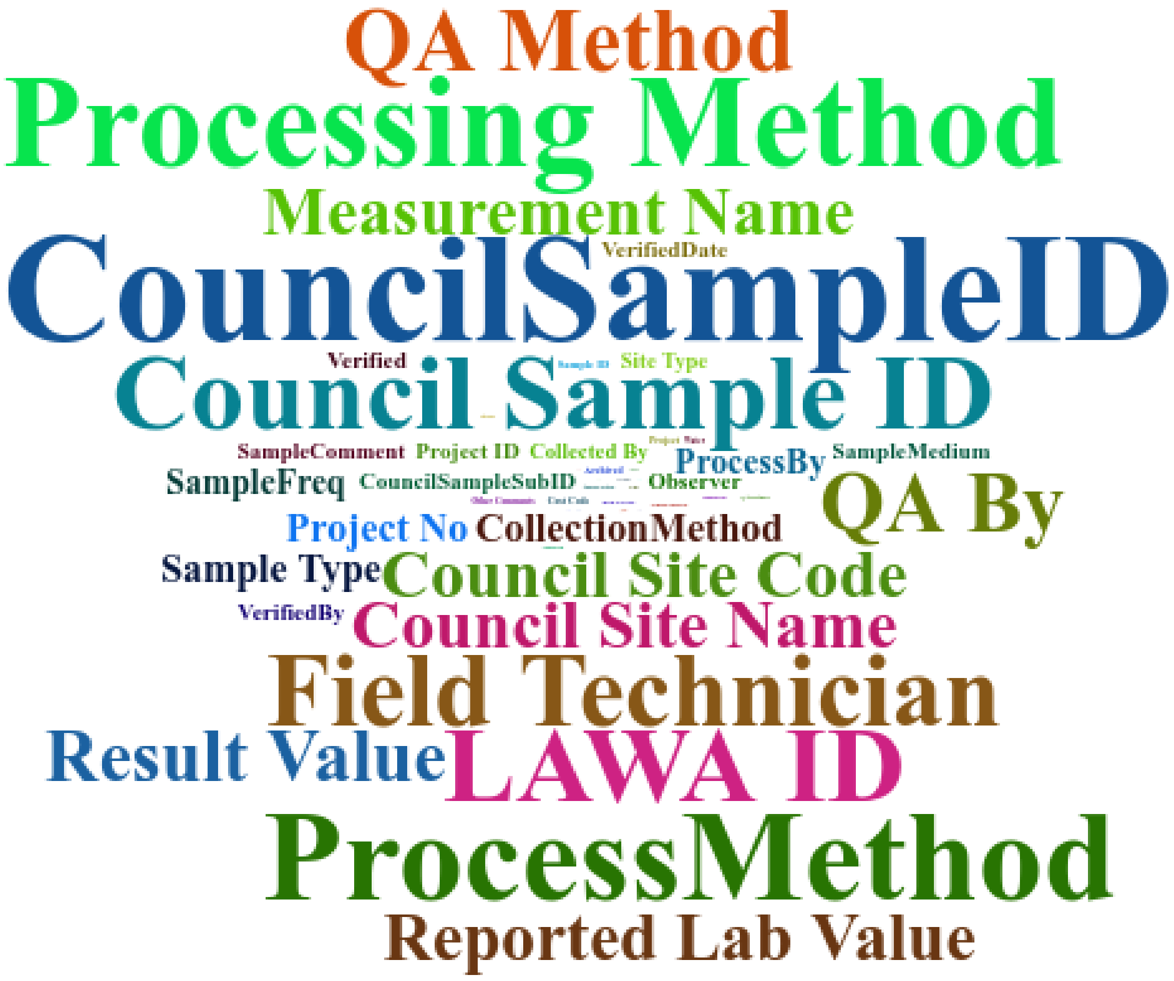
Agency	Rivers	Lakes	Macro	Type of data server
AC	http://aklc.hydrotel.co.nz:8080/KiWIS/KiWIS?datasource=3&Procedure=Sample.Results.LAWA&service=SOS&version=2.0.0&request=GetObservation&temporalfilter=om:phenomenonTime,P25Y/2022-01-01	http://aklc.hydrotel.co.nz:8080/KiWIS/KiWIS?datasource=3&Procedure=Sample.Results.LAWA&Service=SOS&version=2.0.0&request=getObservation&temporalfilter=om:phenomenonTime		KiWIS
BOPRC	http://sos.boprc.govt.nz/service?service=SOS&version=2.0.0&request=GetObservation&temporalfilter=om:phenomenonTime,P15Y/2022-01-01	http://sos.boprc.govt.nz/service?service=SOS&version=2.0.0&request=GetObservation&&temporalfilter=om:phenomenonTime,P15Y/2022-01-01	http://sos.boprc.govt.nz/service?service=SOS&version=2.0.0&request=GetObservation&&temporalfilter=om:phenomenonTime,P15Y/2022-06-01	SOS
ECAN	http://wateruse.ecan.govt.nz/wqlawa.hts?service=Hilltop&Agency=LAWA&request=GetData&From=2004-01-01&To=2022-01-01	http://wateruse.ecan.govt.nz/wqlawa.hts?service=Hilltop&Agency=LAWA&request=GetData&From=2004-01-01&To=2022-01-01	http://wateruse.ecan.govt.nz/wqlawa.hts?service=Hilltop&Agency=LAWA&request=GetData&From=2004-01-01&To=2022-06-01	Hilltop
ES	http://odp.es.govt.nz/WQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://odp.es.govt.nz/SOEFreshwater.hts?service=Hilltop&request=GetData&From=2006-01-01&To=2022-01-01	http://odp.es.govt.nz/MI.hts?service=Hilltop&request=GetData&From=2006-01-01&To=2022-06-01	

Agency	Rivers	Lakes	Macro	Type of data server
GDC	http://hilltop.gdc.govt.nz/data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01		http://hilltop.gdc.govt.nz/data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-06-01	
GWRC	http://hilltop.gw.govt.nz/Data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://hilltop.gw.govt.nz/Data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://hilltop.gw.govt.nz/Data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	Hilltop
HBRC	https://data.hbrc.govt.nz/Envirodata/EMARDiscreteGood.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	https://data.hbrc.govt.nz/Envirodata/EMARDiscreteGood.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	https://data.hbrc.govt.nz/Envirodata/EMARDiscreteGood.hts?service=Hilltop&request=GetData&From=1990-01-01&To=2021-06-01	Hilltop
HRC	https://tsdata.horizons.govt.nz/boo.hts?service=Hilltop&agency=LAWA&request=GetData&From=2004-01-01&To=2022-01-01&ShowQuality=Yes	http://tsdata.horizons.govt.nz/boo.hts?service=Hilltop&agency=LAWA&request=GetData&From=1/1/2004&To=1/1/2025&ShowQuality=Yes	https://tsdata.horizons.govt.nz/boo.hts?service=Hilltop&agency=LAWA&request=GetData&From=1/1/2004&To=1/1/2025&ShowQuality=Yes	Hilltop
MDC	http://hydro.marlborough.govt.nz/LAWA_WQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://hydro.marlborough.govt.nz/LAWA_LWQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2021-01-01	http://hydro.marlborough.govt.nz/LAWA_WQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-06-01	Hilltop
NCC	http://envdata.nelson.govt.nz/data.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01		http://envdata.nelson.govt.nz/data.hts?service=Hilltop&request=GetData&From=1990-01-01&To=2022-06-01	Hilltop

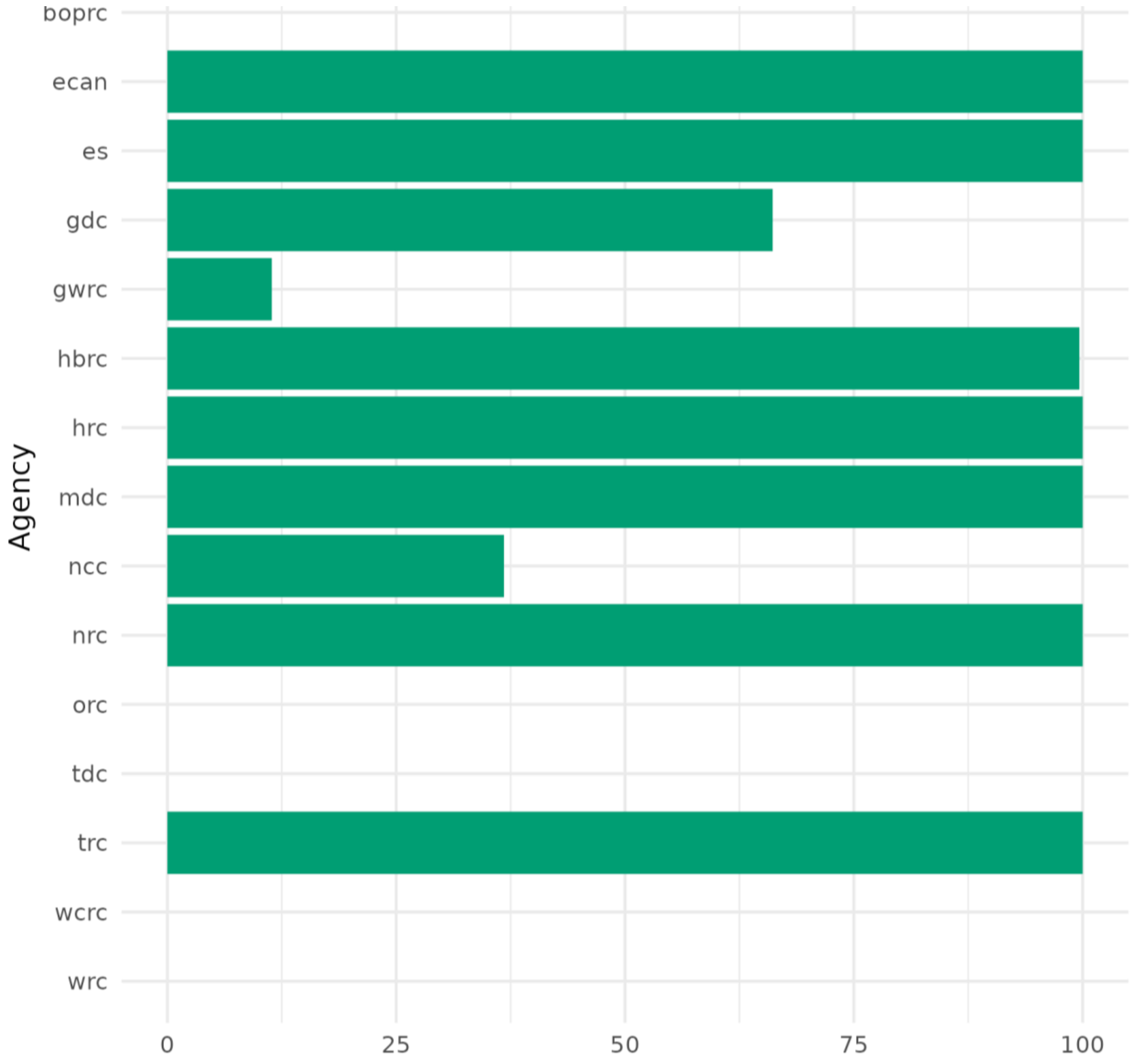
Agency	Rivers	Lakes	Macro	Type of data server
NRC	http://hilltop.nrc.govt.nz/SOEFinalArchive.hts?service=Hilltop&request=GetData&agency=LAWA&From=2004-01-01&To=2022-01-01	http://hilltop.nrc.govt.nz/SOEFinalArchive.hts?service=Hilltop&request=GetData&agency=LAWA&From=2004-01-01&To=2022-01-01	http://hilltop.nrc.govt.nz/SOEMacroinvertebrates.hts?service=Hilltop&request=GetData&From=1999-01-01&To=2022-06-01	Hilltop
ORC	http://gisdata.orc.govt.nz/hilltop/ORCWQ.hts?service=Hilltop&request=GetData&agency=LAWA&From=2004-01-01&To=2022-01-01	http://gisdata.orc.govt.nz/hilltop/ORCWQ.hts?service=Hilltop&request=GetData&agency=LAWA&From=2004-01-01&To=2022-01-01	http://gisdata.orc.govt.nz/hilltop/WQGlobal.hts?service=Hilltop&request=GetData&agency=LAWA&From=1990-01-01&To=2022-06-01	Hilltop
TDC	http://envdata.tasman.govt.nz/WaterQuality.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01		http://envdata.tasman.govt.nz/Invertebrates.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	Hilltop
TRC	https://extranet.trc.govt.nz/getdata/LAWA_river_WQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	https://extranet.trc.govt.nz/getdata/LAWA_lake_WQ.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	https://extranet.trc.govt.nz/getdata/LAWA_bio.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	Hilltop
WCRC	http://hilltop.wcrc.govt.nz/wq.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://hilltop.wcrc.govt.nz/wq.hts?service=Hilltop&request=GetData&From=2004-01-01&To=2022-01-01	http://hilltop.wcrc.govt.nz/wq.hts?service=Hilltop&request=GetData&Site=HRK000085&From=2004-01-01&To=2022-06-01	Hilltop

Agency	Rivers	Lakes	Macro	Type of data server
WRC	http://envdata.waikatoregion.govt.nz:8080/KiWIS/KiWIS?datasource=0&service=SOS&agency=LAWA&version=2.0&request=GetObservation&procedure=RERIMP.Sample.Results.P&temporalfilter=om:phenomenonTime,2004-01-01/2022-01-01	http://envdata.waikatoregion.govt.nz:8080/KiWIS/KiWIS?datasource=0&service=SOS&agency=LAWA&version=2.0&request=GetObservation&procedure=LWQ.Sample.Results.P&temporalfilter=om:phenomenonTime2004-01-01/2022-01-01	http://envdata.waikatoregion.govt.nz:8080/KiWIS/KiWIS?datasource=0&service=SOS&version=2.0&request=GetObservation&procedure=Cmd.P&temporalfilter=om:phenomenonTime,P30Y	KiWIS

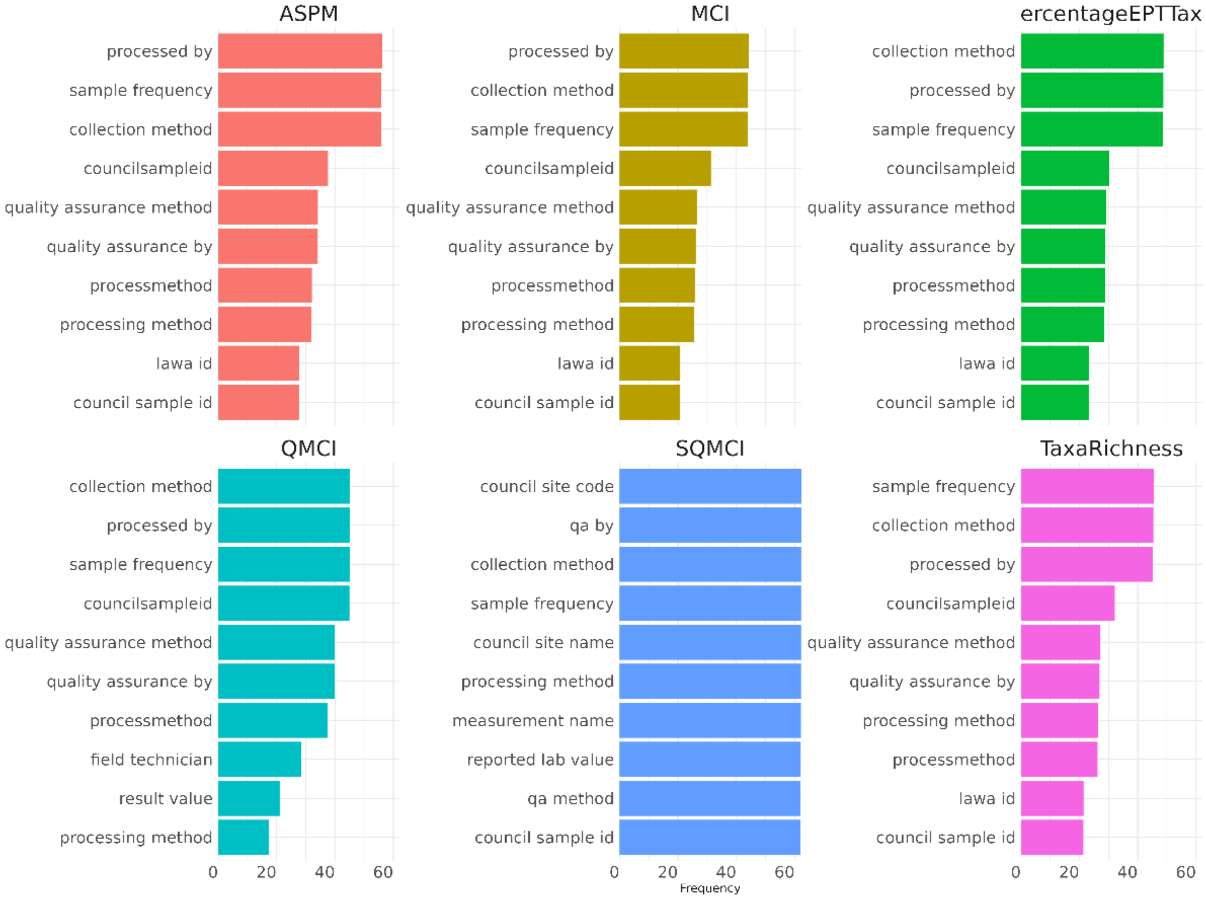
Appendix 3. Word cloud with all the different metadata variables for macroinvertebrates data. The word size is based on the frequency that each metadata variable was used.



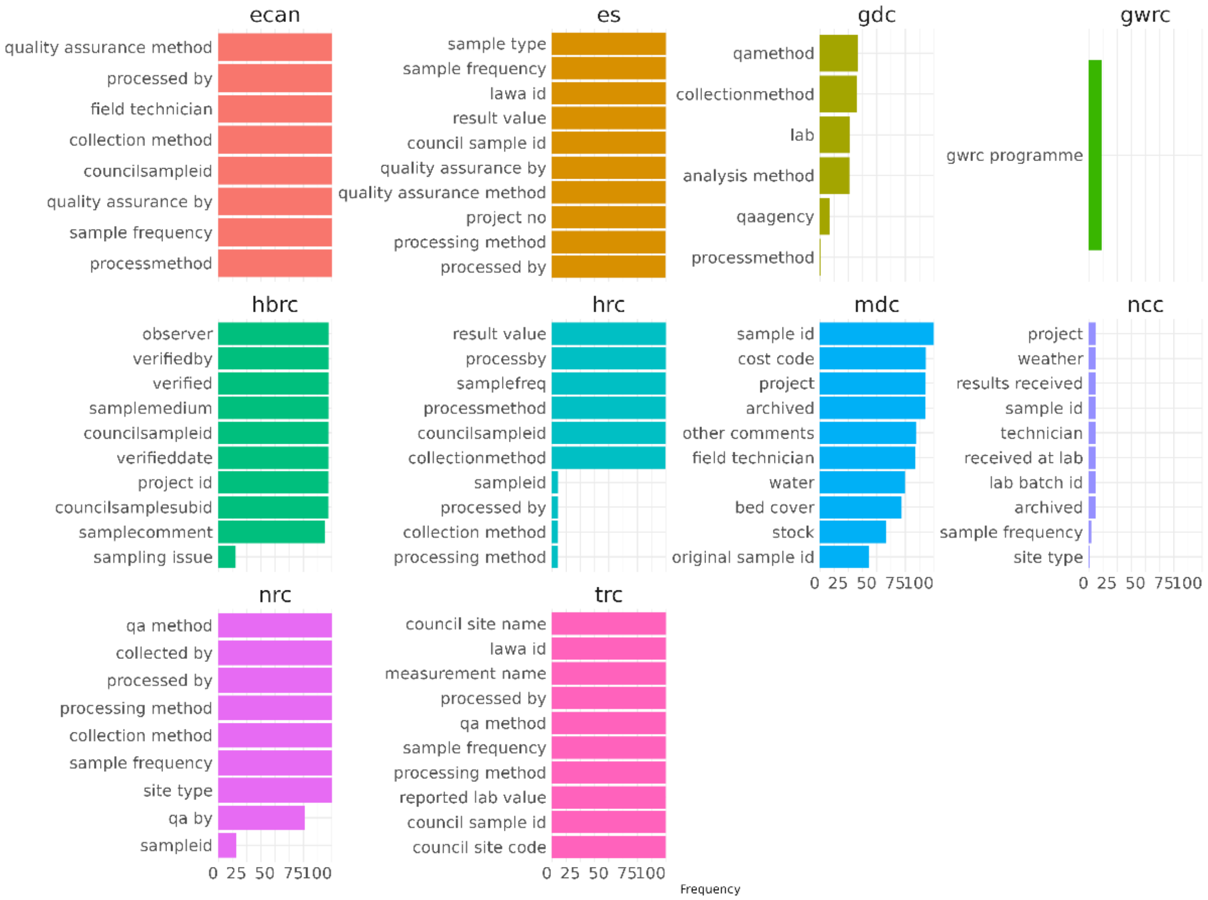
Appendix 4. Proportion of macroinvertebrate data with associated metadata per council. Only councils that provide their data through a data server were included in this analysis; for those not showing in this figure (e.g. Auckland Regional Council – AC), data were not available via a data server request. For an explanation of abbreviations, see Glossary.



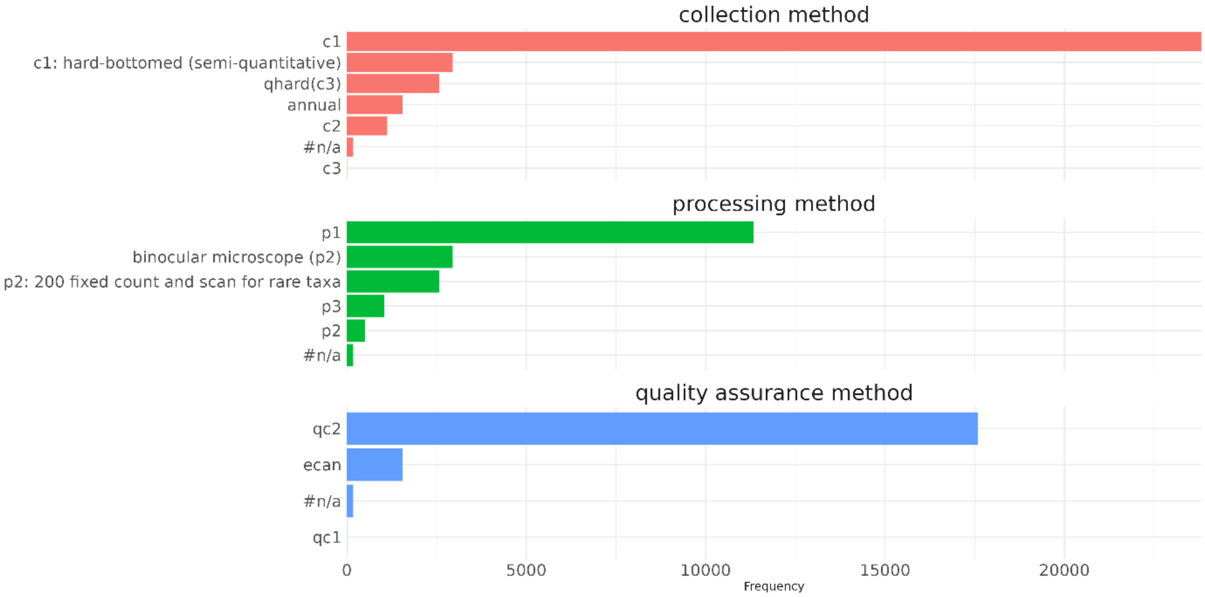
Appendix 5. Proportion of records associated to the 10 most frequent metadata variables used by each parameter for the macroinvertebrates module. For an explanation of abbreviations and terms, see Glossary.



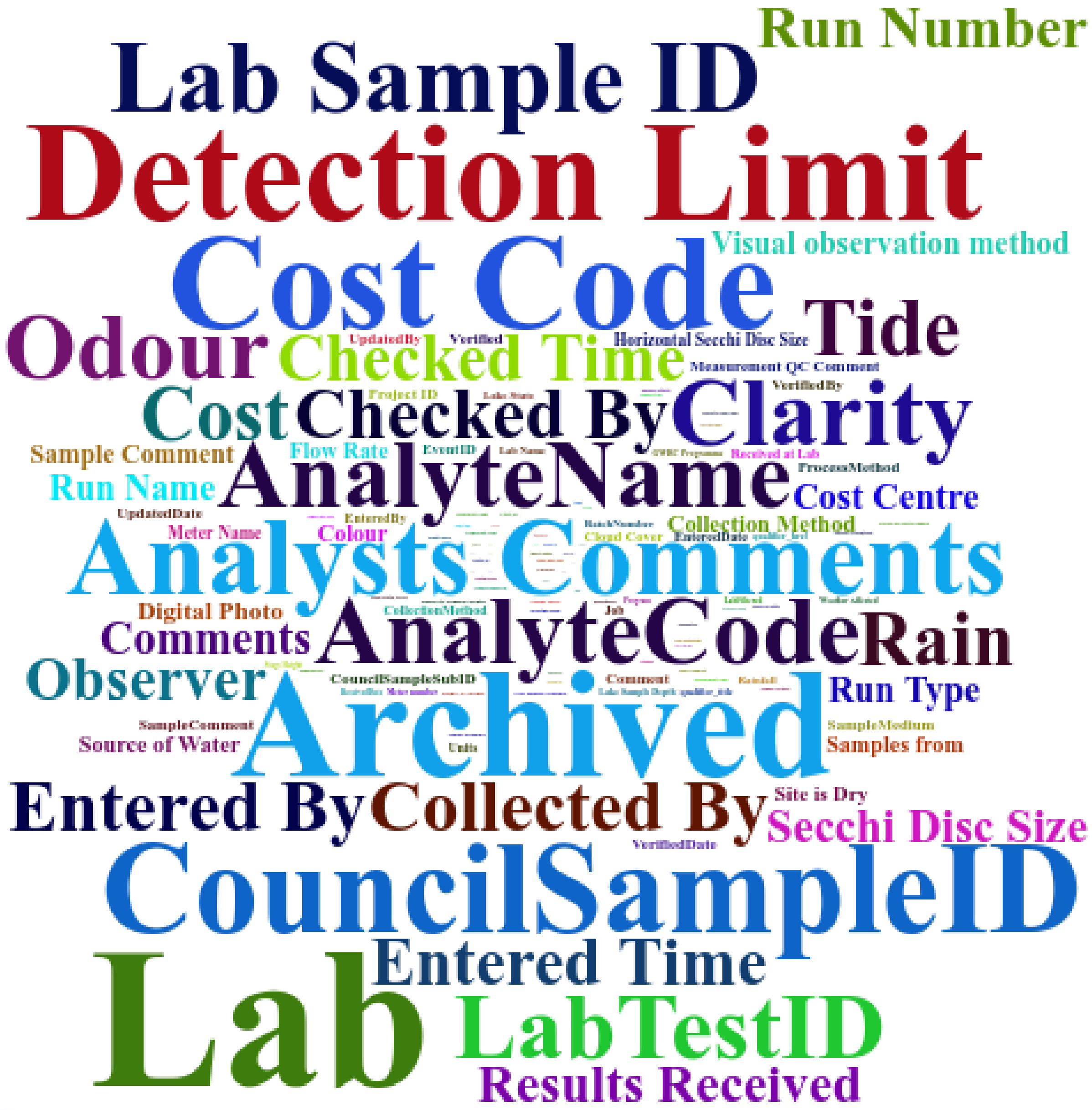
Appendix 6. Proportion of records associated to the 10 most frequent metadata variables used by each council for the macroinvertebrates module. For an explanation of abbreviations, see Glossary.



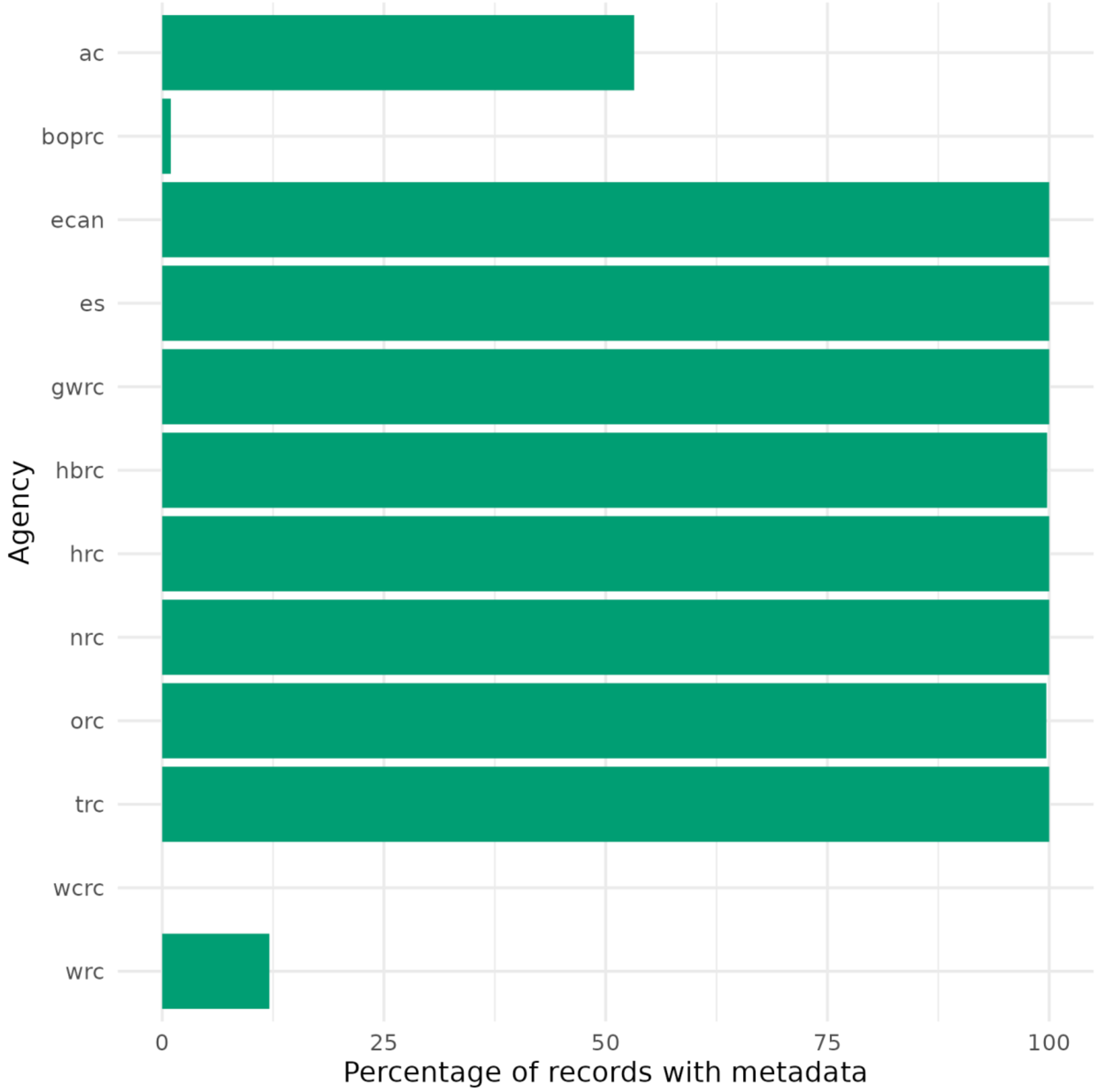
Appendix 7. Number of records associated to the most frequent metadata values for three metadata variables relevant to the laboratory analysis of the data for the macroinvertebrates module.



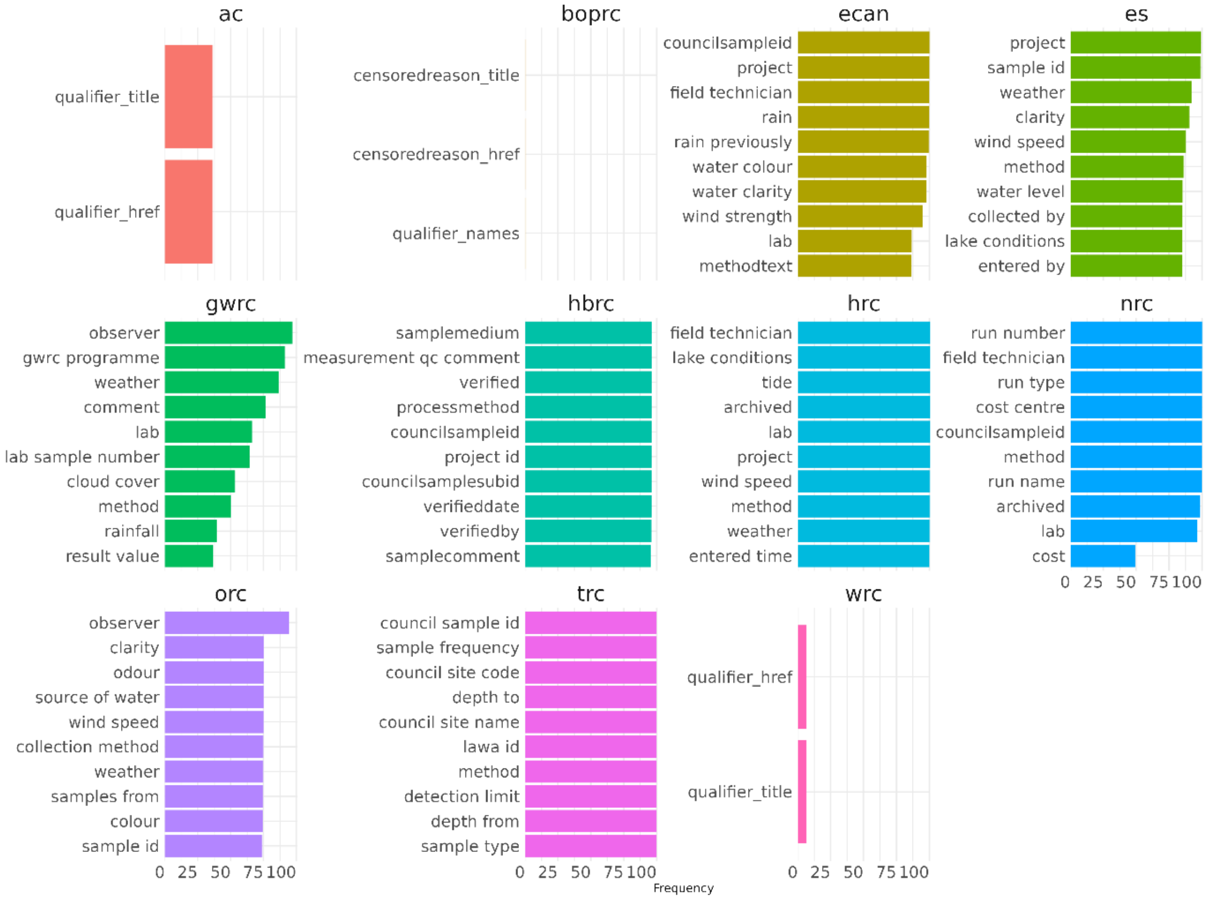
Appendix 8. Word cloud with all the different metadata variables for lakes data. The word size is based on the frequency that each metadata variable was used.



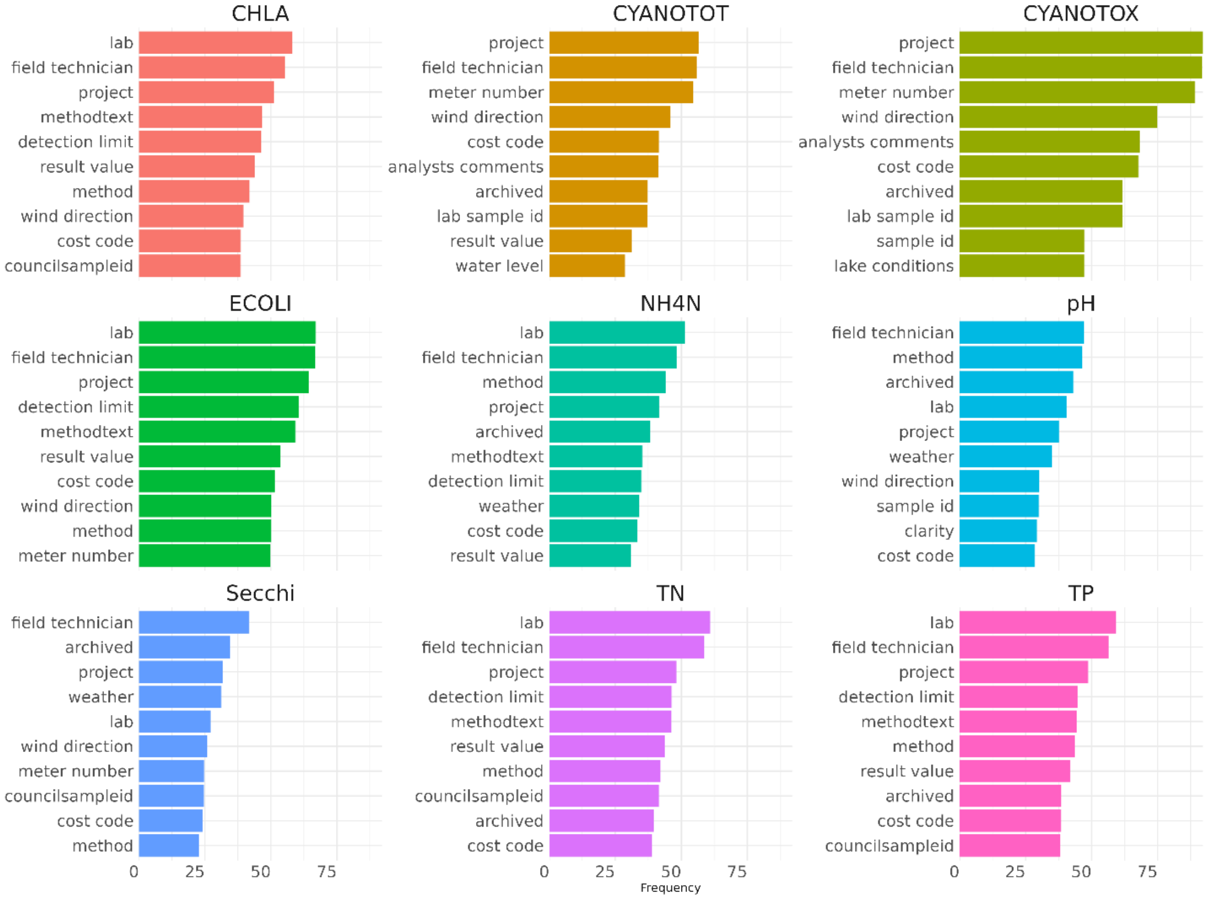
Appendix 9. Proportion of lakes data with associated metadata per council. Only councils that provide their data through a data server were included in this analysis. For an explanation of abbreviations, see Glossary.



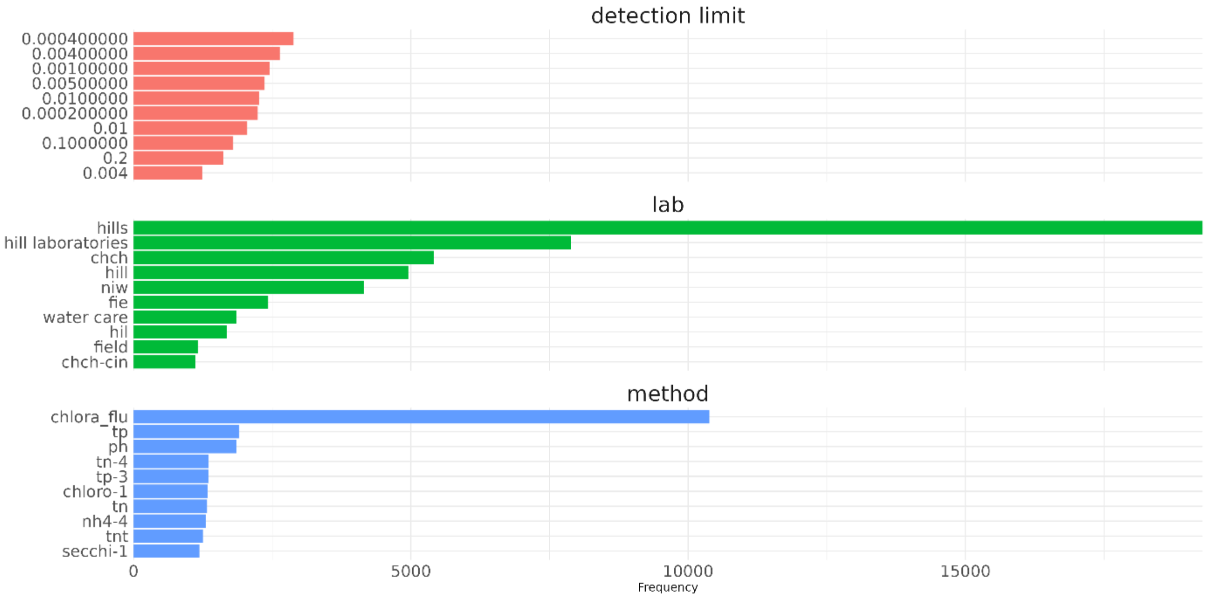
Appendix 10. Proportion of records associated to the 10 most frequent metadata variables used by each council for the lakes module. For an explanation of abbreviations, see Glossary.



Appendix 11. Proportion of records associated to the 10 most frequent metadata variables used by each parameter for the lakes module. For an explanation of abbreviations and terms, see Glossary.



Appendix 12. Number of records associated to the most frequent metadata values for three metadata variables relevant to the laboratory analysis of the data for the lakes module.



9. REFERENCES

[DES] Department of Environment and Science. 2018. Monitoring and sampling manual: environmental protection (water) policy. Brisbane: Department of Environment and Science Government. https://environment.des.qld.gov.au/__data/assets/pdf_file/0031/89914/monitoring-sampling-manual-2018.pdf

[NEMS]. National Environmental Monitoring Standards. 2013. National quality code schema. Version 1.0. https://bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com/public/Documents/NEMS-Quality-Code-Schema-v1.0.pdf

[NEMS]. National Environmental Monitoring Standards. 2016. National quality code schema. Version 2.0. https://bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com/public/Documents/NEMS-Quality-Code-Schema-v2.0.pdf

[NEMS]. National Environmental Monitoring Standards. 2019a. Water quality. Part 2 of 4: sampling, measuring, processing and archiving of discrete river water quality data. Version 1.0.0. https://bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com/public/Documents/NEMS-Water-Quality-Part-2-Sampling-Measuring-Processing-and-Archiving-of-Discrete-River-Water-Quality-Data-v1.0.0.pdf

[NEMS] National Environmental Monitoring Standards. 2019b. Water quality. Part 3 of 4: sampling, measuring, processing and archiving of discrete lake water quality data. Version 1.0.0. https://bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com/public/Documents/NEMS-Water-Quality-Part-3-Sampling-Measuring-Processing-and-Archiving-of-Discrete-Lake-Water-Quality-Data-v1.0.0.pdf

[NEMS] National Environmental Monitoring Standards. 2022. Macroinvertebrates. Collection and processing of macroinvertebrate samples from rivers and streams. Version 1.0.0. https://bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com/public/Documents/Macroinvertebrates-v1.0.0.pdf

[NEMS] National Environmental Monitoring Standards. 2023. Long-term strategy. NEMS-Strategy-May-2023.pdf (bucketeer-54c224c2-e505-4a32-a387-75720cbeb257.s3.amazonaws.com)

R Core Team. 2022. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org

RStudio Team 2022. RStudio: integrated development environment for R. Boston (MA): RStudio. http://www.rstudio.com