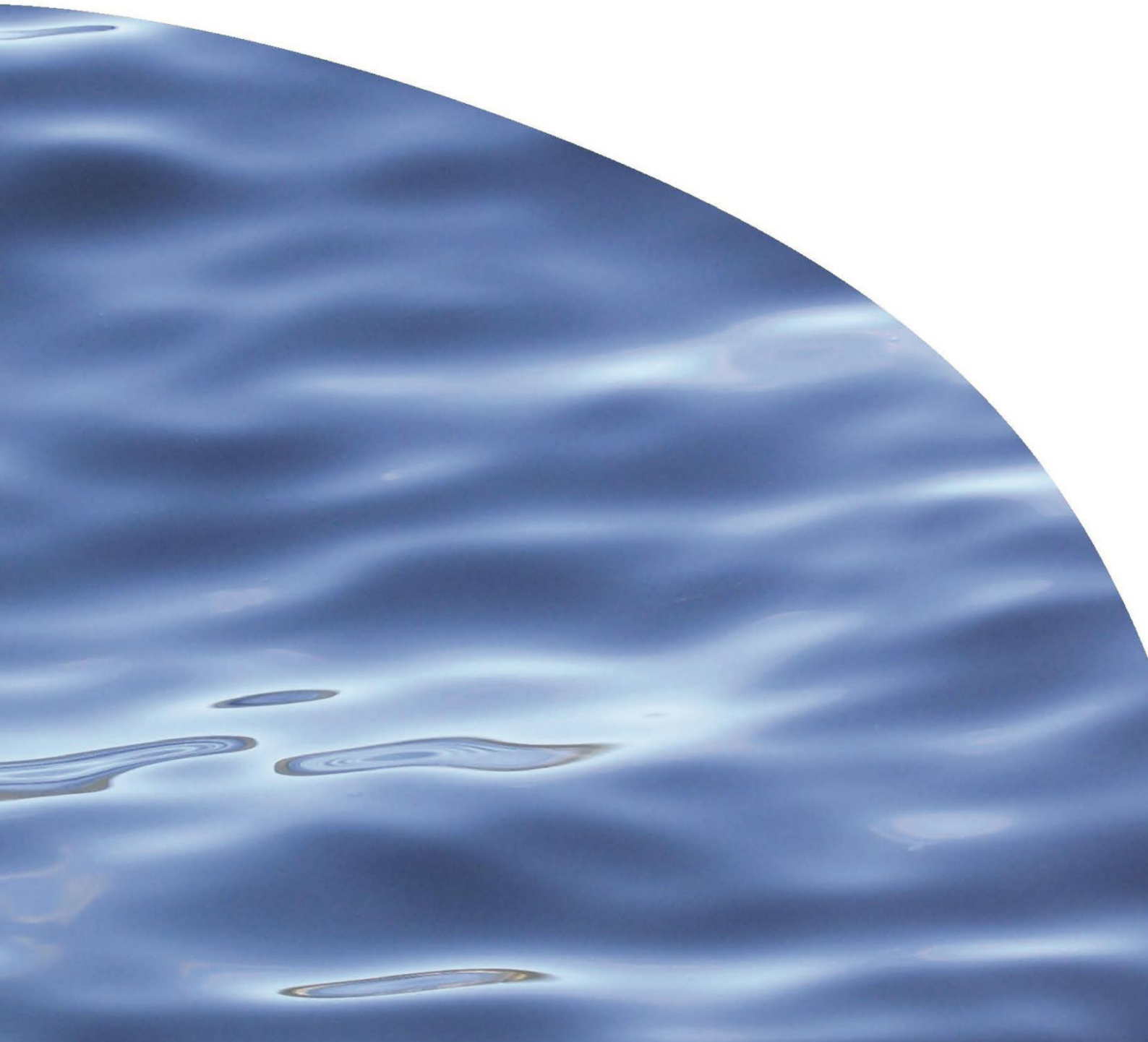
CAWTHRON
The power of science®

REPORT NO. 3073

MACROINVERTEBRATE METRICS FOR THE NATIONAL POLICY STATEMENT FOR FRESHWATER MANAGEMENT



MACROINVERTEBRATE METRICS FOR THE NATIONAL POLICY STATEMENT FOR FRESHWATER MANAGEMENT

JOANNE CLAPCOTT, ANNIKA WAGENHOFF, MARTIN NEALE[1] RICHARD STOREY[2], BRIAN SMITH[2], RUSSELL DEATH[3], JON HARDING[4], CHRISTOPH MATTHAEI[5], JOHN QUINN[2], KEVIN COLLIER[6], JAVIER ATALAH, ERIC GOODWIN, HAYDEN RABEL, JAMES MACKMAN, ROGER YOUNG

[1] Martin Jenkins Limited
[2] National Institute of Water and Atmospheric Research
[3] Massey University
[4] University of Canterbury
[5] University of Otago
[6] University of Waikato
All others Cawthron Institute

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:
Dr. Charles P Hawkins
(Utah State University)

RELEASED BY:
Chris Cornelisen

ISSUE DATE: 31 October 2017

RECOMMENDED CITATION: Clapcott J, Wagenhoff A, Neale M, Storey R, Smith B, Death R, Harding J, Matthaei C, Quinn J, Collier K, Atalah J, Goodwin E, Rabel H, Mackman J, Young R 2017. Macroinvertebrate metrics for the National Policy Statement for Freshwater Management. Prepared for the Ministry for the Environment. Cawthron Report No. 3073. 139 p. plus appendices.

EXECUTIVE SUMMARY

This document is the final written output from the Ministry for the Environment funded project on benthic macroinvertebrate indicators of ecosystem health (Contract 21630). The project was designed to address a recognised need to include macroinvertebrates in the National Policy Statement for Freshwater Management (NPS-FM) 2014. Benthic macroinvertebrates are used worldwide as sub-indicators of stream ecosystem health as they respond to human pressures, are taxonomically diverse and easy to sample. In New Zealand, the macroinvertebrate metrics that are most commonly used in environmental reporting include variants of the Macroinvertebrate Community Index (MCI) and of the three insect orders Ephemeroptera, Plecoptera and Trichoptera (EPT) taxa. During the progress of the project, monitoring of MCI became compulsory in an amendment of the NPS-FM (2017). The MCI is responsive to multiple stressors, but not all stressors, and as such provides a good indicator of the overall condition of the macroinvertebrate component of stream ecosystem health. However, the MCI is not diagnostic and cannot inform specific management decisions on resource use. Subsequently, this project includes research supporting the development of new stressor-specific macroinvertebrate metrics (e.g. sediment, nutrients) as well as value-specific macroinvertebrate metrics (i.e. Ecosystem Health as defined in the NPS-FM[1]).

The primary objectives of this study were to define the quantitative relationship between macroinvertebrate metrics (new and existing) and human stressors and to explore the connection between macroinvertebrate metrics and the Ecosystem Health (EH) value. In doing so, the applicability of using these metrics to assess the EH value in the NPS-FM was tested. To address the research objectives the following tasks were undertaken:

- collation of existing data and calculation of existing metrics including updating the macroinvertebrate species traits database (Section 2)
- proof of concept of new stressor-specific metrics (Section 3)
- exploration of a multivariate approach to assessing EH (Section 4)
- characterisation of the quantitative link between metrics and stressors (Section 5)
- development of a framework to include macroinvertebrate metrics in the NPS-FM to assess the Ecosystem Health value (Section 6).

Collation of existing data and calculation of existing metrics

Three national datasets were collated in this study for the calculation of existing metrics and development of new stressor-specific and value-specific metrics. The first dataset comprised macroinvertebrate community data from regional councils and National River Water Quality Network sites and was compiled by a parallel project on sediment attributes. Data were standardised to a common taxonomic level before we calculated 26 existing macroinvertebrate metrics based on taxa relative abundance or taxa presence-absence. Metrics based on taxa density were excluded due the inconsistency in sampling methods. In

[1] 'In a healthy freshwater ecosystem ecological processes are maintained, there is a range and diversity of indigenous flora and fauna, and there is resilience to change.' (NPS-FM 2014).

parallel, we updated the New Zealand macroinvertebrate trait database (maintained by NIWA) by incorporating new knowledge from published studies and expert opinion to determine the trait 'profile' of each taxon. These profiles were used to calculate a trait matrix, or trait composition as opposed to taxonomic composition, for each sample in this national dataset. The trait composition for any given site included 59 trait modalities. We also calculated 13 trait diversity indices.

The second dataset comprised macroinvertebrate community data and associated physicochemical variables (deposited sediment, suspended sediment, nutrients, periphyton, temperature, dissolved oxygen) collected during research projects. These data were compiled from published and unpublished research by the project team. The third dataset comprised macroinvertebrate community data from reference sites and included a subset of data from both previous datasets as well as additional research data. All data were standardised to a common taxonomic level.

Development towards new stressor-specific metrics

We explored the development of stressor-specific metrics using the second research dataset. Stressor-specific metrics are useful tools for managing ecosystem health because they can help identify the main cause(s) of stream degradation at specific sites and/or at a regional/national level. Once causes and limiting factors have been identified, stressor-specific metrics could be used to track restoration success or effectiveness of regional policies over time with respect to management of a specific stressor. Methods for metric development included systematic and non-systematic reviews of the scientific literature, the assignment of tolerance values to taxa based on the expert opinion of team members, and the assignment of tolerance values using a gradient forest approach. The relative abundance of taxa was analysed in relation to major stressor gradients in New Zealand. The gradient forest outputs were used to develop taxa tolerance scores which were then used to calculate new stressor-specific metrics. Data limitations meant the gradient forest approach was only suitable for the development of sediment-specific and enrichment-specific metrics, which included 20 new metrics.

Exploration of a multivariate approach to assessing EH

We explored a multivariate predictive model approach to assessing the components of ecosystem health represented by macroinvertebrate communities with the third reference dataset. Multivariate predictive models are used globally to provide a reference condition-based assessment of stream communities, but a national macroinvertebrate model is yet to be developed for New Zealand. Based on the biological classification of sites, we made predictions by constructing a River Invertebrate Prediction and Classification System (RIVPACS) reference condition-type model. The predictive accuracy of the biological classification model was high and similar to that observed for other multivariate models developed overseas. We also tested a multivariate model based on a stream typology—the Freshwater Ecosystems of New Zealand (FENZ) classification of sites, which performed equally well. This work represents the early development of a predictive model that could provide a robust basis for assessing macroinvertebrate communities in New Zealand rivers.

Quantifying the link between metrics and stressors

Defining the quantitative link between macroinvertebrate metrics and manageable stressors determines which metrics can be used in a stressor-specific context to help set limits to meet freshwater objectives and to monitor progress towards those objectives. We explored the relationship between all 110 metrics, taxonomic and trait composition, and measures of catchment condition and proximate stressors using gradient forest and general linear model analyses of the first two datasets combined. The gradient forest analyses were used to identify which metrics had the largest relative effect sizes and consistent response shapes in relation to catchment-scale land use and in-channel proximate stressors. The linear models were used to quantify the relationship between metrics and drivers and identify how much independent variance could be assigned to specific drivers. This latter analysis identified that some of the newly developed stressor-specific metrics could be considered truly stressor-specific, whereas existing tolerance metrics such as the MCI are responsive to multiple stressors and hence good indicators of the multiple impact pathways of land use on stream ecosystem health.

Framework to assess the macroinvertebrate component of stream ecosystem health

Finally, a framework for the inclusion of macroinvertebrate metrics in the NPS-FM to assess the Ecosystem Health value was developed. Following an international approach, we identified a combination of metrics that represent the key properties of ecosystem health (EH) including organisation/composition, richness/diversity, functional aspects and tolerance. The combination of metrics that best distinguished reference from non-reference sites were identified using logistic regression. In the table below, 4 key metrics in bold contribute equally to an overall multi-metric score for a macroinvertebrate sub-index of stream ecosystem health. The remaining 6 functional and tolerance metrics provide diagnostic tools for further assessing the pathways through which degradation, or conversely rehabilitation, is occurring. Additional diagnostic metrics can be incorporated into the framework as they become available.

Ecosystem health property	**Organisation / composition**	**Richness / Diversity**	**Functional aspects**	**Tolerance**
Key metric(s)	**%EPT richness**	**EPT richness**	**CPI1** (One reproductive cycle per individual)	**MCI hb**
Diagnostic metrics			Aduolar (Adult or larvae aquatic stages)	Sediment: Sed MCI Submerged (Submerged ovipositor) Crawlers (Epibenthic) Enrichment: Chl MCI Lowflex (Low body flexibility)

Multi-metric scores range from 0 to 1 and were grouped into 4 management classes reflecting deviation from reference state:

A – at or similar to natural state
B – low deviation from natural state
C – high deviation from natural state
D – substantial deviation from natural state.

The spatial variation in the multi-metric scores showed that 24% of all State of the Environment sites were at or similar to reference state on average in the last 3 years and 31% of all sites showed substantial deviation from reference. Multi-metric scores increased in response to native vegetation cover and decreased in response to pastoral heavy cover. The inter-annual variation in multi-metric scores was twice as high on average at non-reference sites compared to reference sites (< 85% native vegetation catchments). Of the component metrics, MCI_hb (hard bottom) had the lowest inter-annual variation and was the most likely to correctly distinguish reference from non-reference at the national scale.

Conclusion and recommendations

The results of this study support the recent inclusion of the MCI in the NPS-FM 2014 (amended 2017); it is a sensitive indicator of the multiple stressor effects on macroinvertebrates resulting from dominant land uses in New Zealand, and can be used to distinguish the ecosystem health of streams at a national scale. It is, however, only one indicator and cannot be used to identify specific stressors nor inform catchment and in-stream resource use. We have provided proof-of-concept for the development of new stressor-specific metrics. In particular, we have developed metrics that discriminate between sediment and enrichment effects and which could be used as diagnostic tools to inform resource use and restoration priorities. We recommend scientific validation of these metrics so that they can be adopted nationally. We have further shown the potential of a multivariate, or RIVPACS styled approach to assessing the health of the macroinvertebrate community. We recommend further development of a national multivariate model, which could in turn be used alongside metrics. Finally, we recommend that 4 key metrics be used to calculate a multi-metric which can be used to assess the macroinvertebrate component of stream ecosystem health. The deviation from reference state can be used to assign management classes. We suggest the deviation from reference condition approach be further strengthened by the development of models predicting stream-type specific reference benchmarks for core metrics, because the current approach used to assign management classes may show bias against some stream types. An additional 6 metrics can be used to diagnose the cause of stream degradation and we hope, to monitor improvement over time in response to stream rehabilitation. These diagnostic metrics can be added to over time as new stressor-specific metrics are developed. We suggest priority be placed on collecting new macroinvertebrate data paired to temperature and flow data to aid in the development and validation of temperature and flow stressor-specific metrics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

GLOSSARY

AMDI	Acid Mine Drainage Index	Acronym
ASPM	Average score per metric	Acronym
C	Celsius	Unit
DIN	Dissolved inorganic nitrogen	Acronym
DOC	Department of Conservation	Acronym
DRP	Dissolved reactive phosphorus	Acronym
EH	Ecosystem health	Acronym
EPA	Environmental Protection Authority	Acronym
EPT	Ephemeroptera, Plecoptera and Trichoptera	Acronym
FENZ	Freshwater Ecosystems of New Zealand	Acronym
GIS	Geographical Information System	Acronym
IBI	Index of biological integrity	Acronym
km	Kilometre	Unit
LIFENZ	New Zealand Lotic Index for Flow Evaluation	Acronym
m	Metre or metres	Unit
m^3/s	Cubic metres per second	Unit
MCI	Macroinvertebrate community index	Acronym
MfE	Ministry for the Environment	Acronym
mm	Millimetres	Unit
N	Nitrogen	Abbreviation
NH4	Ammonium	Abbreviation
NIWA	National Institute of Water and Atmospheric Research	Acronym
NOF	National Objective Framework	Acronym
NPS-FM	National Policy Statement for Freshwater Management	Acronym
NRWQN	National River Water Quality Network	Acronym
NZ	New Zealand	Acronym
NZReach	A unique identifier available for every stream segment (length of stream between tributary junctions) in the REC digital river network	Abbreviation
O/E	Ratio of observed to expected	Acronym
QMCI	Quantitative Macroinvertebrate Community Index	Acronym
RBP	Rapid Bioassessment Protocol	Acronym
REC	River Environment Classification	Acronym
RF	Random forest	Acronym
RIVPACS	River Invertebrate Prediction and Classification System	Acronym
SD	Standard deviation	Acronym
SIS	Suspendable inorganic sediment	Acronym
SQMCI	Semi Quantitative Macroinvertebrate Community Index	Acronym
SoE	State of the Environment	Acronym
TN	Total nitrogen	Acronym
TON	Total organic nitrogen	Acronym
TP	Total phosphorus	Acronym
taxon	Taxonomic group of any rank, such as a species, family, or class	Acronym
UCI	Urban community index	Acronym

1. INTRODUCTION

This document is the final written output from the Ministry for the Environment (MfE)-funded project on benthic macroinvertebrate indicators of ecosystem health (Contract 21630). The project addressed a recognised need to include macroinvertebrates in the National Policy Statement for Freshwater Management (NPS-FM) 2014. It includes research into the development of stressor-specific macroinvertebrate metrics (e.g. sediment, nutrients) as well as value-specific macroinvertebrate metrics (e.g. ecosystem health). In this report, initial metric development is described along with statistical development linking metrics to stressors and a framework for how to include macroinvertebrates in the NPS-FM is recommended.

1.1. Benthic macroinvertebrates

Benthic macroinvertebrates are used worldwide as indicators of stream ecosystem health as they respond to human pressures, are taxonomically diverse and easy to sample. In New Zealand, the macroinvertebrate metrics that are most commonly used in environmental reporting include variants of the Macroinvertebrate Community Index (MCI) and taxa of the three insect orders Ephemeroptera, Plecoptera and Trichoptera (EPT). These two core metrics are sensitive in their response to human impacts on streams, but they are not stressor-specific or indeed sensitive to all stressors[2] (Clapcott & Goodwin 2014; Collier et al. 2014). For example, the MCI which was originally developed to measure the effect of organic pollution on the macroinvertebrate community responds to multiple stressors (Figure 1) and this makes it difficult to determine the primary stressor and subsequently, how to manage for a change in the MCI value. This is an issue with biotic indices worldwide.

[2] 'Stressors' are variables such as nutrients and sediments and other pollutants whose concentrations are exacerbated by human activities.

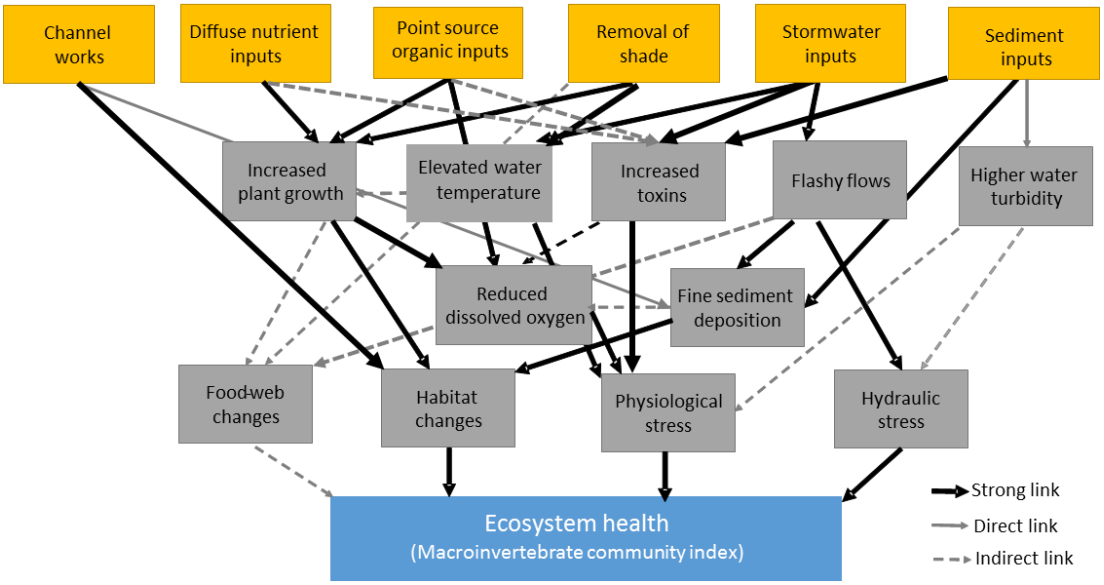


Figure 1. Pathways (hypothesised links) by which various stressors (orange boxes) influence the macroinvertebrate community index (MCI) indicator of ecosystem health. Adapted from (Collier et al. 2014).

1.2. New Zealand policy context

The National Policy Statement for Freshwater Management (NPS-FM) 2014 (amended 2017) requires regional councils, through their regional plans, to set freshwater objectives that provide for freshwater values, and to set resource use limits and management actions to achieve those objectives (Figure 2). The NPS-FM includes the National Objectives Framework (NOF), which defines attributes that assist regional councils to set freshwater objectives (i.e. numeric) and justifiable policies (including limits) for achieving these.

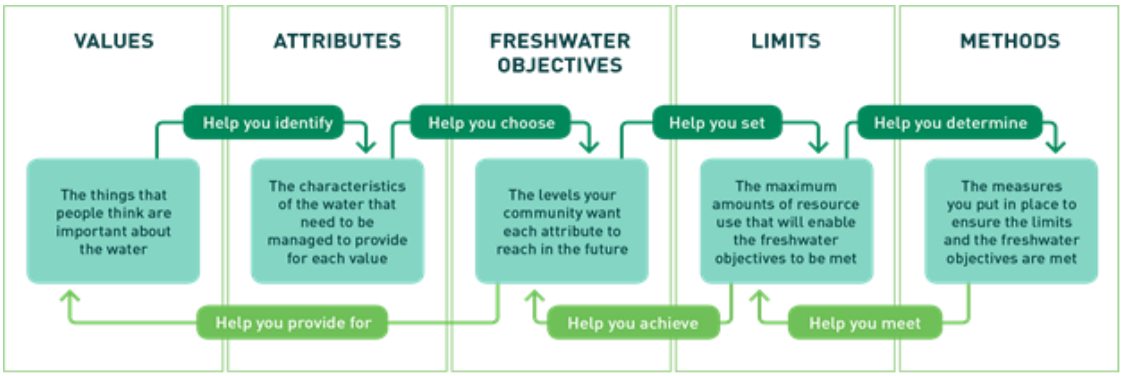


Figure 2. The links between freshwater values, attributes, objectives, limits and management methods on the National Objectives Framework (Ministry for the Environment 2015).

The NPS-FM includes two compulsory national values that apply to all streams — Ecosystem Health and Human Health. Regional councils are required to develop monitoring plans that monitor progress towards, and the achievement of, objectives associated with these values. Specifically, Policy Objective CB1(aa) requires councils to develop a monitoring plan that must at least include the monitoring of macroinvertebrate communities. Under Policy Objective CB3 every regional council must be:

a) using the Macroinvertebrate Community Index;

b) establishing methods under Policy CB2 to respond to a Macroinvertebrate Community Index score below 80, or a declining trend; and

c) ensuring that methods:

   i. investigate the causes of declining trends or the Macroinvertebrate Community Index score below 80;

   ii. seek to halt declining trends; and

   iii. seek to improve on a Macroinvertebrate Community Index score if it is below 80, unless this is caused by naturally occurring processes, pest or unwanted organism, or by infrastructure listed.

Previously, the MCI had been excluded as an attribute because it is linked to land use through a complex chain of causality which makes isolating the role of specific stressors difficult, and hence the setting of limits on catchment and water resource use problematic (Clapcott & Goodwin 2014). However, those conclusions were based on limited analysis of regional datasets. The recent inclusion of MCI in the 2017 amended NPS-FM recognises the value of MCI as an indicator of overall stream health. Nonetheless, councils still require macroinvertebrate metrics to investigate the cause of declining trends in MCI scores and to contribute to a holistic assessment of Ecosystem Health, even if they are not included as attributes within the NOF.

1.3. Scope of this study

- The primary objectives of this study were to define the quantitative relationship between macroinvertebrate metrics (new and existing) and human stressors and to explore the connection between macroinvertebrate metrics and the Ecosystem Health (EH) value. In doing so, the applicability of using macroinvertebrate metrics to measure the EH value in the NPS-FM was tested. To address the research objectives the following tasks were undertaken:
- collation of existing data and calculation of existing metrics including updating the macroinvertebrate species traits database (Section 2)
- proof of concept of new stressor-specific metrics (Section 3)

- exploration of a multivariate approach to assessing EH (Section 4)
- characterisation of the quantitative link between metrics and stressors (Section 5)
- development of a framework to include macroinvertebrate metrics in the NPS-FM to assess the Ecosystem Health value (Section 6)

1.4. Author contributions

This project was a team effort with 15 contributing authors. All authors contributed to this report via recommended changes and comments on a draft version, they also attended project meetings and workshops where they contributed expert opinion on stressor-specific metric development and frameworks for multi-metrics. Primary authors of the individual sections were: Section 1 (Joanne), Section 2 (Annika, Richard), Section 3 (Annika, James), Section 4 (Martin), Section 5 (Annika, Javier), Section 6 (Joanne) and Section 7 (Joanne).

2. DATASETS AND THE CALCULATION OF EXISTING MACROINVERTEBRATE METRICS

2.1. Overview

This section describes the collation of existing data and the calculation of existing metrics as well as an update of the macroinvertebrate species traits database. The aim of this task was to create an up-to-date database on macroinvertebrate data and stressor information from across the country. It required the compiling of State of the Environment (SoE), National River Water Quality Network and research macroinvertebrate data, and accompanying metadata and physicochemical variables where measured at corresponding sites. It required taxa auditing and data harmonisation to a consistent taxonomic level. The compiled dataset was then used to calculate a range of macroinvertebrate metrics and traits (described in this section) as well as further project tasks (described in following sections).

2.2. Collation of existing data

Benthic macroinvertebrate data and associated land use and stressor data were compiled from a range of sources. The primary source was the national SoE network. Other sources included the National River Water Quality Network (NRWQN) and published and unpublished research datasets. Each dataset is described below.

2.2.1. National datasets

Macroinvertebrates

The national macroinvertebrate dataset consisting of benthic macroinvertebrate data collected at SoE sites by regional or unitary councils, and data collected by NIWA at NRWQN sites. This dataset was compiled by Martin Unwin (NIWA) and provided through a parallel MfE project concerned with developing sediment attributes for future incorporation into policy (Depree et al. 2017).

The dataset comprised 15,508 samples collected from 1966 stream sites spread throughout New Zealand during the period from 1990 to 2016. In the majority of cases a single site was sampled on a yearly basis (up to 23 years), with few cases where a site has been sampled more than once per year. Each site was labelled by NZReach; a unique identifier of each stream segment in the River Environment Classification digital network (Snelder et al. 2004).

Macroinvertebrate samples were collected using quantitative or semi-quantitative methods as outlined in Stark et al. (2001). Data were given in counts or coded abundances depending on which of the standard processing protocols (Stark et al. 2001) had been used. According to these protocols, coded abundance classes are

specified as Rare (1-4), Common (5-19), Abundant (20-99), Very Abundant (100-499) and Very Very Abundant (500+ individuals). The area of streambed sampled was not provided preventing the calculation of macroinvertebrate densities.

Taxa were typically identified to the lowest practical taxonomic levels using New Zealand's standard key (Winterbourn et al. 2006 or earlier editions). For example, EPT taxa were mostly identified to species level if possible. The dataset was checked for taxonomic revisions that occurred over the data timeframe, and where possible taxon names were adjusted accordingly. Some revisions were not able to be accounted for, e.g. where a genus was reclassified into two separate genera. In such cases it is not possible to go back and update the data. However, overall taxonomic changes have been few and this would be a minor source of error in the dataset. Two versions of the final dataset were prepared, one where the original taxonomic resolution was kept and one where the taxonomic level was aggregated to the level required for calculation of MCI type metrics (typically genus level).

Water quality

Water quality data were extracted from the SoE data available on the LAWA website on 5 May 2017 (https://www.lawa.org.nz/). Of interest were data on nutrient concentrations (total phosphorus (TP), total nitrogen (TN), dissolved reactive phosphorus (DRP), ammonium (NH4) and total organic nitrogen (TON)) and suspended fine sediment (turbidity and black disc clarity) as nutrients and sediment are the two main stressors affecting macroinvertebrate communities.

Periphyton

Periphyton data were obtained from SoE and NRWQN sites, representing a combined total of 1,031 sampling locations. Metrics were calculated from data collected using standard methods (Biggs & Kilroy 2000). Four metrics were sufficiently widely available to establish a sound basis for subsequent analyses: chlorophyll-*a* (mg/m^2); % cover of long filaments; % cover of thick mats; and % total cover (equivalent to the sum of long filament and thick mats). At least one of these variables was available for 5,810 samples from 2000 to 2016, representing a total of 1,031 sites.

Deposited and suspended fine sediment

Deposited sediment data were compiled as part of a parallel MfE project on the development of sediment attributes for the NOF (Depree et al. 2017). Three metrics were sufficiently widely available to be used in future analyses: % cover fine sediment (< 2 mm) in a reach (riffle, run and pool habitat) assessed from the stream bank, % cover fine sediment (< 2 mm) in run habitat assessed instream; and suspendable inorganic sediment (SIS, g/m^2) in run habitat. Data were collected using standard methods (SAM1, SAM2 and SAM4 described in Clapcott et al. (2011)). Percentage cover of fine sediment assessments according to SAM1 or SAM2 have shown to be highly correlated while only the instream % fine sediment cover (SAM2)

was correlated with SIS (SAM4) (Clapcott et al. 2011). At least one of these three sediment metrics was available for 3,274 samples from 1999 to 2016, representing a total of 702 sites. The same dataset also contained data on suspended fine sediment (total suspended solids, turbidity, and water clarity assessed with a black disc).

2.2.2. Research dataset

We collated 26 macroinvertebrate research datasets[3] that contained data on stressors including deposited fine sediment and/or on nutrients (nitrogen, phosphorus) and/or periphyton (chlorophyll-*a*, total percent cover) collected during the period from 1987 to 2017. The final research dataset contained 1,861 samples collected from 973 sites spread throughout New Zealand. GPS coordinates were used to retrieve the respective NZReach number using a spatial join in ArcGIS. The accuracy of spatial joins was visually checked.

A mixture of quantitative and semi-quantitative methods were used to collect and then process macroinvertebrate samples, predominately following standard protocols (Stark et al. 2001) to provide taxa densities or estimates of taxa relative abundances from counts or from coded abundances. The common denominator of abundance data was relative abundance, hence all data were expressed as proportions. For coded abundance data, the lower boundary of the respective abundance classes was used to derive proportional data. The taxonomic resolution was mainly at lowest practical taxonomic levels and old taxon names were updated in the same way as for the national dataset (see Section 2.1.1). Two versions of the final dataset were prepared: one where the original taxonomic resolution was kept and one where the taxonomic level was aggregated to coarser taxonomic resolution required for calculation of MCI type metrics.

2.2.3. Reference dataset

We assembled a dataset of macroinvertebrate taxonomic data from 538 reference sites defined by the presence of greater than 80% native vegetation in upstream catchments. The dataset included sites from the SoE data compilation described previously (Section 2.2.1) supplemented by unpublished research datasets, primarily from Russell Death (and Massey University colleagues), Richard Storey (NIWA), Dave West (DOC) and Lyndsay Chadderton (formerly DOC).

The macroinvertebrate data varied in terms of taxonomic resolution and abundance recording as they had been collected for a range of purposes; therefore we aggregated abundance data to the taxa level required for calculation of MCI-type metrics (typically genus level). This dataset was compiled for the purpose of exploring a multivariate approach to assessing EH (see Section 4).

[3] References for the studies and an outline of research data are given in Appendix 3 (Tables A3.2 and A3.3, respectively).

2.3. Overview of metrics considered and data requirements

We considered a wide range of macroinvertebrate metrics that are in use and/or have been developed in New Zealand (Table 1). The majority of the metrics considered require either information on the presence only of taxa or on the relative abundances of taxa requiring some method of enumeration. For this type of information, it is sufficient to collect macroinvertebrates semi-quantitatively, e.g. with a kick-net. A smaller amount of the metrics considered require quantitative data, e.g. collected with a Surber sampler. However, the exact area of the streambed sampled when Surber samples had been taken was not given and likely varied depending on the size of the sampling device and the number of replicates taken. This prevented the calculation of metrics that require densities, i.e. number of individuals per area streambed sampled. While about 50% of the samples in the dataset were collected quantitatively, information on the total area of the streambed sampled was not given, preventing calculation of those metrics.

Sample processing in the laboratory also determines what type of metric can be calculated. Determining a list of the taxa present is the simplest way of processing the data. However, for all samples in this dataset, processing involved counting of the individuals within each taxon, which can be done in different ways. A rapid method to enumerate taxa is to place them in abundance categories providing coded abundance data which was done for 30% of the samples (standard protocol P1 described in Stark et al. 2001).

Another important property of the macroinvertebrate data for metric calculation is the taxonomic resolution. Metrics that are based on tolerance values require the taxonomic resolution that is at least to the level for which tolerance values have been developed. Diversity/richness metrics were calculated using the taxonomic resolution that was provided by the councils or the research dataset providers. Inconsistencies in the resolution provided by different taxonomists (or sample processors) could affect comparability of the diversity/richness metrics among sites. However, as each individual dataset (i.e. from each council or researcher) is usually processed by a single person in a given year and covers wide stressor gradients, inconsistencies in the resolution can be considered noise and are unlikely to affect stressor-response relationships that this work is concerned with.

All indices were calculated in the statistical programme R. Finally, we did not calculate the number or percentage of exotic species because of a general absence of non-indigenous taxa in New Zealand benthic macroinvertebrate samples.

Table 1. Macroinvertebrate metrics considered along with data requirement. EPT metrics denoted with * indicates that Hydoptilidae were excluded from calculation, metrics denoted with † were not calculated (see text for explanation), hb and sb subscripts for MCI indices relate to use of tolerance values developed by Stark (1985, and updates) and Stark and Maxted (2007) for either hard-bottomed or soft-bottomed streams respectively, hb2 subscript for MCI indices relate to use of tolerance values developed by Greenwood et al. (2015). A description of the calculation of these metrics as well as formula is in Section 2.3.

Metric	Description	Data requirement
MCI_{hb}	Macroinvertebrate Community Index (hard-bottomed)	presence-absence
$QMCI_{hb}$	Quantitative MCI (hard-bottomed)	relative abundance
$SQMCI_{hb}$	Semi-quantitative MCI (hard-bottomed)	coded relative abundance
MCI_{sb}	MCI (soft-bottomed)	presence-absence
$QMCI_{sb}$	Quantitative MCI (soft-bottomed)	relative abundance
$SQMCI_{sb}$	Semi-quantitative MCI (soft-bottomed)	coded relative abundance
MCI_{hb2}	MCI (Greenwood tolerance scores)	presence-absence
$QMCI_{hb2}$	Quantitative MCI (Greenwood tolerance scores)	relative abundance
$SQMCI_{hb2}$	Semi-quantitative MCI (Greenwood tolerance scores)	coded relative abundance
Density†	Number of individuals per streambed area	density
Biomass†	Weight of individuals per streambed area	density
Productivity	Change in biomass per year	presence-absence
EPT taxon richness	Number of Ephemeroptera, Plecoptera, Trichoptera taxa	presence-absence
% EPT taxon richness	Percentage of EPT taxa	presence-absence
% EPT abundance	Percentage abundance of EPT	relative abundance
EPT taxon richness	Number of EPT taxa excluding Hydroptilidae	presence-absence
% EPT taxon richness	Percentage of EPT taxa excluding Hydroptilidae	presence-absence
% EPT abundance	Percentage abundance of EPT excluding Hydroptilidae	relative abundance
% exotics†	Number of non-indigenous taxa	presence-absence
LIFENZ	New Zealand Lotic Index for Flow Evaluation	coded relative abundance
LIFENZ_W	weighted LIFENZ	coded relative abundance
AMDI	Acid Mine Drainage Index	presence-absence
UCI	Urban Community Index	presence-absence
QUCI	Quantitative UCI	relative abundance
ASPM	Average Score Per Metric	relative abundance
$ASPM_Q$	ASPM calculated from site scores scaled using percentiles	relative abundance
Taxon richness	Number of taxa	presence-absence
Pielou's evenness index	Distribution of individuals among taxa	relative abundance
Simpson's diversity index	Based on the number and abundance of taxa	relative abundance
Functional diversity indices (multiple)	Based on number and abundance of traits	relative abundance

2.4. Details on the calculation of the various metrics

2.4.1. Macroinvertebrate Community Index (MCI) metrics

Tolerance values

The MCI is a metric which was developed to indicate organic and nutrient pollution and is similar to others used around the world (e.g. the Biotic Index developed by Hilsenhoff (1977)). Its development followed the approach of the British National Water Council's Biological Monitoring Working Party and is described in Stark (1985). The data used to develop the MCI were from a study of riffles in stony streams that formed part of the Taranaki Ringplain Water Resources Survey (Stark 1985). First, stream sites were divided into three pollution classes (unpolluted streams, slightly to moderately polluted streams, or grossly polluted streams) based on professional judgement. Next, tolerance values were assigned by first calculating for each taxon (usually genus) the mean relative abundance across sites within each of the three pollutional classes, and secondly, calculating the weighted average of the three mean percentages using the weighting factors 10, 5 and 1 for the least polluted, intermediately polluted and most polluted stream classes, respectively. For about 70% of the taxa, tolerance values were directly obtained by this numerical procedure although some scores were modified by ±1 point based on professional judgement. For the remaining 30% of the taxa, which were infrequently present in samples, professional judgement was used to assign tolerance scores. Finally, resulting MCI site scores were compared across the three pollution classes originally identified.

The MCI was developed for stony (hard-bottomed) streams and based on a regional data set (Taranaki); however, the index soon found widespread use throughout New Zealand including for soft-bottomed streams. For soft-bottomed streams, the MCI site scores were generally found to be lower than those in hard-bottomed streams (despite similar water quality) triggering the development of tolerance values specifically for soft-bottomed streams. These tolerance values were developed using an objective iterative rank correlation procedure (Stark & Maxted 2007b). This procedure was adopted from Chessman (2003) who used it to update tolerance values for Australian stream macroinvertebrates to be used in the biotic index called SIGNAL2. Briefly, Chessman (2003) describes the method as follows. First, rank correlation coefficients, expressed as a proportion of the maximum correlation mathematically possible, were calculated between the original biotic index scores and abundances of each taxon across all samples in the dataset. These adjusted correlation coefficients were used to assign tolerance values to the taxa. The taxon with the highest correlation was assigned a score of 10 and the one with the lowest a score of 1. All other taxa were scaled between these values proportional to their correlations. This process was repeated several times until the tolerance values stabilised.

Since the development of the MCI, tolerance values for hard-bottomed streams have been progressively updated and last published together with tolerance values for soft-

bottomed streams by Stark and Maxted (2007a). We used a further updated table of tolerance values by John Stark (personal communication, January 2017) provided in Appendix 1 (Table A1.1). We calculated both the hard-bottomed and soft-bottomed versions of the MCI for all samples, notated by subscript 'hb' or 'sb'.

Recently, tolerance values were re-calculated using a national-scale macroinvertebrate dataset and Chessman's (2003) iterative rank correlation procedure described above to produce tolerance values for as many freshwater macroinvertebrate taxa as possible (Greenwood et al. 2015). The taxonomic resolution was mostly similar to that used for the original MCI, but with some exceptions. There were insufficient data from soft-bottomed streams, hence tolerance values were developed for 240 taxa for use in hard-bottomed streams (Appendix 1, Table A1.2). To distinguish MCI site scores calculated with those updated tolerance values, this MCI is labelled with subscript 'hb2'.

MCI variants

There are three types of MCI metrics that depend on the data type provided and that were calculated both with the tolerance values provided by John Stark (Table A1.1) and those provided by Greenwood et al. (2015, Table A1.2). The Macroinvertebrate Community Index (MCI) is calculated based on presence data as follows

$$\mathrm{MCI} = \frac{\sum_{i=1}^{i=S} a_i}{S} \mathrm{x}\ 20$$

where S = the total number of scoring taxa in the sample, and a_i = the tolerance value for the ith taxon. The scaling factor of 20 has been added to distinguish between MCI site scores and site scores of the MCI variants taking into account taxon counts (QMCI and SQMCI).

The Quantitative Macroinvertebrate Community Index (QMCI) is calculated based on count data derived from quantitative samples (e.g. Surber) as follows

$$\mathrm{QMCI} = \sum_{i=1}^{i=s} \frac{(n_i \ \mathrm{x}\ a_i)}{N}$$

where S = the total number of taxa in the sample, n_i = the abundance for the i^{th} scoring taxon, a_i = the tolerance value for the i^{th} taxon (see Appendices 1 and 2) and N = the total abundance of the scoring taxa for the entire sample. Note that calculation of the QMCI does not need the information of the streambed area sampled.

The SQMCI is calculated similarly to the QMCI but typically based on coded abundance data (assigned to the R = Rare, C = Common, A = Abundant, VA =Very Abundant and VVA = Very Very Abundant classes) from semi-quantitative samples (e.g. kicknet) as follows

$$\text{SQMCI} = \sum_{i=1}^{i=S} \frac{(n_i \text{ x } a_i)}{N}$$

where S = the total number of scoring taxa in the sample, n_i = the coded abundance for the i^{th} scoring taxon, a_i = the tolerance value for the i^{th} taxon and N = the total of the coded abundances of the scoring taxa for the entire sample. The coded abundance is the lower boundary of the respective abundance classes assigned to each taxon based on the counts: R (1-4), C (5-19), A (20-99), VA (100-499) and VVA (500+).

The SQMCI was developed to provide an index that takes into account information on the relative abundances of taxa, as the QMCI, but that is more cost-effective to determine. Semi-quantitative sampling in the field is faster than quantitative sampling, however the biggest time savings can be achieved by using faster enumeration methods in the laboratory. Use of coded abundance categories is usually the fastest method followed a fixed count of 200 individuals plus a scan for rare taxa and finally a full count. Hence, the SQMCI is typically calculated from coded abundance data although it can also be calculated from fixed count or full count data. The difference between the SQMCI calculated from coded abundance or relative abundance data has shown to be minor (Stark 1998). Finally, the SQMCI has shown to provide a similar assessment to the QMCI with less than 40% of the effort (Stark 1998). Correspondence between the QMCI and MCI has also been tested and site scores were found to rank sites similarly (Wright-Stow & Winterbourn 2003).

2.4.2. EPT metrics

Species within the insect orders Ephemeroptera (E), Plecoptera (P) and Trichoptera (T) are generally sensitive to pollution and hence are used worldwide as indicator taxa in stream health metrics. In New Zealand, there are two genera, *Oxyethira* and *Paroxyethira*, both belonging to the family Hydroptilidae whose dominant species are relatively tolerant to pollution and hence often get excluded from metric calculations. We calculated both versions, with and without the Hydroptilidae family, using the original taxonomic resolution provided.

We calculated EPT taxon richness and %EPT taxon richness, both of which can be calculated from presence-absence data. EPT taxon richness is the number of taxa belonging to the orders Ephemeroptera, Plecoptera or Trichoptera, and %EPT taxon richness is the percentage of *taxa* that belong to the orders Ephemeroptera,

Plecoptera or Trichoptera. Relative abundance data also allows calculation of %EPT abundance which is the percentage of *individuals* that belong to the orders Ephemeroptera, Plecoptera or Trichoptera. We calculated these for count data as well as for the coded abundance data using the same conversion of coded abundance categories to count data as for calculation of MCI metrics (see Section 2.4.1).

2.4.3. New Zealand Lotic Index for Flow Evaluation

The New Zealand Lotic Index for Flow Evaluation (LIFENZ) was developed for the purpose of determining impacts of river flow alterations on aquatic ecosystems (Greenwood et al. 2016). Similarly to the UK-based LIFE metric (Extence et al. 1999), taxon indicator values represent preference for one of four flow (water velocity) category ('specific', 'moderate', 'general', 'ultra-general'). Values were assigned to 193 aquatic macroinvertebrate taxa using professional judgement. The LIFENZ site score is calculated as follows

$$\text{LIFENZ} = \frac{\sum_{i=1}^{i=S} \text{fs}}{S}$$

where S = the total number of scoring taxa in the sample, fs = the flow score which is determined for each scoring taxon based on its assigned flow category and coded abundance (table 4 in Greenwood et al. 2016).

The LIFENZ_W variant down-weights the scores if the taxon has a general compared to a more specific velocity preference. Flow preference categories are presented alongside the flow categories for all taxa in Greenwood et al. (2016, Appendix A).

We calculated the LIFENZ and LIFENZ_W only for samples that have been processed using the protocol generating coded abundances. The taxonomic resolution required is similar to that of the MCI.

2.4.4. Acid Mine Drainage Index

The Acid Mine Drainage Index (AMDI) was developed for assessing coal mining impacts in New Zealand streams by associating water chemistry and benthic macroinvertebrate community data collected from 91 sites on the West Coast of the South Island (Gray & Harding 2012). Indicator values for 57 taxa were calculated using weighted averaging and range from 0 to 10. The AMDI site score is calculated from presence-absence data as follows

$$\text{AMDI} = \left\{ \frac{\sum_{i=1}^{i=S} a_i}{S} * \log_{10} S \right\} * 10$$

where S = the total number of scoring taxa in the sample and a_i = the tolerance value for the *i*th taxon. A richness multiplier ($\log_{10} S$) has been incorporated into the formula as richness is known to be a useful indicator of acid mine drainage impacts (Gray & Harding 2012). A scaling factor of 10 results in site scores ranging from 0 (severely impacted) to 100 (unimpacted). Sites can then be categorised as 'severely impacted', 'impacted' or 'unimpacted'. The taxonomic resolution required is similar to that of the MCI.

2.4.5. Urban Community Index

The Urban Community Index (UCI) was developed to assess the degradation of physical habitat quality in urban streams. Suren et al. (1998) analysed a macroinvertebrate dataset using Canonical Correspondence Analysis to retrieve ordinal scores of each site based on their macroinvertebrate communities. Sites with similar taxon compositions have similar ordinal scores. The method also calculates taxon ordination scores, which were used as tolerance values to assign to 91 taxa (Appendix 1, Table A1.3).

The UCI site score is calculated using presence-absence data, in the same fashion as the MCI, as follows:

$$\mathrm{UCI} = \frac{\sum_{i=1}^{i=S} a_i}{S} \mathrm{x}\ 20$$

where *S* = the total number of scoring taxa in the sample, and α_i = the tolerance value for the *i*th taxon.

The Quantitative Urban Community Index (QUCI) takes into account the relative abundances (counts) of each taxon and is calculated, in the same fashion as the QMCI, as follows

$$\mathrm{QUCI} = \sum_{i=1}^{i=S} \frac{(n_i \ \mathrm{x}\ a_i)}{N}$$

where S = the total number of scoring taxa in the sample, n_i = the abundance for the *i*th scoring taxon, a_i = the tolerance value for the *i*th taxon (see Table A1.3 and N = the total abundance of the scoring taxa for the entire sample. Note that while the name of the QUCI suggests that quantitative data are needed, count data from a semi-quantitative sample (unknown area of streambed area) is sufficient. The taxonomic resolution required is similar to that of the MCI.

2.4.6. Average Score Per Metric

The Average Score Per Metric (ASPM) is a multi-metric index developed by Collier (2008). It is calculated from three metrics, the MCI (we used the hard-bottomed version), EPT taxon richness and %EPT abundance (both EPT metrics excluding Hydroptilidae), hence requiring counts only of EPT taxa and total numbers. These three metrics were selected from a suite of 17 candidate metrics for their ability to discriminate between reference sites and sites influenced by urbanisation or high levels of pastoral development in the Waikato region (Collier 2008). Metrics were aggregated by firstly scaling (normalising) the observed site scores by the observed maximum value across a set of sample sites resulting in values between 0 and 1 (Collier 2008). Then the mean of these scaled metrics was used to calculate the ASPM.

We calculated the ASPM following Collier (2008), scaling the site scores using the formula

$$x' = [x - x_{min})] / [x_{max} - x_{min}]$$

where x' is the scaled site score, x is the raw site score and x_{min} and x_{max} are the minimum and maximum site scores of the entire national dataset. The range in site scores in our dataset was 0 to 0.84 as the maximum observed site scores were relatively high (MCI_{hb}: 200, EPT taxon richness (excl. Hydroptilidae): 29, % EPT abundance (excl. Hydroptilidae): 100%.

As extreme outliers (at both ends) in the national dataset may affect the ability of the ASPM to discriminate between different levels of land-use impacts, we also calculated the ASPM from scaled scores replacing the minimum and maximum value by the 5th and 95th percentile, respectively. This approach produced ASPM scores of < 0 or > 1 (range: -0.27 to 1.22).

2.4.7. Diversity indices

The most fundamental description of the nature of a macroinvertebrate community is provided by a measure of its diversity: the number of different species of organism and their abundance, generally in terms of individuals, but sometimes biomass. Diversity, or biodiversity, is often assumed by the public to be higher in more pristine environments, although this is often not the case as diversity can be reduced by both high and low disturbance levels. However, to ensure we did not miss any useful index we evaluated several diversity indices to cover both the richness and evenness components of diversity. Taxa richness is the simplest index of species diversity and

we used a simple count of the number of taxa collected[4]. Evenness evaluates how individuals are distributed amongst those species, for example whether one species comprises most of the individuals in a sample. All diversity indices were calculated using the original taxonomic resolution provided as this is the resolution most councils and scientists would use (see discussion on potential biases in Section 2.2).

The equation for Pielou's evenness index is:

$$\text{Pielou's evenness index} = \frac{-\sum_{i=1}^{i=S} p_i \ln p_i}{\ln S}$$

where S = the total number of taxa in the sample and p_i = the relative abundance for the *i*th taxon.

The Simpson's Diversity index is one of the most commonly-used diversity indices (Magurran 2004) and the equation is:

$$D = \sum \left(\frac{n_i(n_i - 1)}{N(N-1)}\right)$$

where n_i = the number of individuals in the ith species and N = the total number of individuals collected.

2.4.8. Functional diversity indices

We calculated a range of functional diversity indices using R package 'FD'. Calculation of these indices uses the trait information described in 2.4. Rao's quadratic entropy (RaoQ), for example, is an index commonly used in the literature for multi-trait functional indices (e.g. Bêche & Resh 2007; Lange et al. 2014). We also calculated Functional Richness (FRic), Functional Divergence (FDiv) and Functional Evenness (FEve). To provide a measure of trophic functional diversity, we also calculated Rao's Q (RaoQ_13) and Functional Evenness (FEve_13) as well as Community Weighted Mean (CWM) for the functional feeding trait only. The CWM index is given for each of the six feeding trait modalities (CWM_13a, CWM_13b, etc). Details on the calculation of the functional diversity indices can be found in the R package 'FD' documentation or in the literature (Bêche & Resh 2007; Casanoves et al. 2011).

[4] Taxa richness, although intuitively simple, suffers from being sensitive to sample effort. Thus the more animals collected, the more taxa are likely to be recorded. To account for this the number of taxa can be corrected for the collected number of individual animals. This is termed rarefied taxa richness (Coleman 1981). However, as count data were not consistently available we only calculated taxa richness.

2.4.9. Productivity

Mean individual lengths and biomasses for each species were determined from the literature and length-biomass regressions (Winterbourn et al. 1989; Towers et al. 1994; Moore 1998; Benke et al. 1999; Baumgärtner & Rothhaupt 2003; Stoffels et al. 2003). Invertebrate biomass (dry weight) was converted to joules following Brey et al. (2010). Annual production for each species was estimated using Brey's (2012) artificial neural network model (internal cross validation $R^2 = 0.801$) with an assumed annual median temperature of 13°C. The estimated production rates were similar to those derived for the same or similar taxa throughout other parts of New Zealand (Hopkins 1976; Huryn 1996; Winterbourn 1996; Huryn 1998; Collier et al. 2004). Annual respiration for each species was estimated using Brey's (2010) artificial neural network model (internal cross validation $R^2 = 0.847$) with an assumed annual median temperature of 13°C. At each site, the average gross macroinvertebrate production and respiration were calculated for the entire samples, averaged per individual and then multiplied by 100 (i.e. joules/100ind/yr).

2.5. Species traits

2.5.1. The potential advantages of using traits

A biological community can be described either by its taxonomic composition or by the composition of the traits present. Species traits are defined as 'a measurable property of an organism, such as body size, longevity, or feeding guild, usually measured in individuals and applied comparatively across species and at broad geographic scales' (McGill et al. 2006 cited in Culp et al. 2011).

Theoretical ecology provides a foundation for understanding how functional trait characteristics are expected to respond to environmental gradients (Townsend & Hildrew 1994; Poff 1997; Statzner et al. 2001). Since Thienemann (1918) and Southwood (1977), habitat has been considered as a trait 'filter' whereby increasing stress filters out unsuitable traits and results in a narrower range of traits present in the invertebrate community. As a result of such filtering, a traits approach has potential for inferring the mechanisms by which the community composition is shaped (Culp et al. 2011). Several studies have shown that the trait composition of benthic macroinvertebrate communities can discriminate among sites differing in level of overall human impact, e.g. upstream vs. downstream of a waste water treatment plant (Charvet et al. 1998) or across a gradient of agricultural impacts (Doledec et al. 1999) or water abstraction (Lange et al. 2014).

Some studies (summarised in Statzner & Beche 2010; also Schuwirth et al. 2015) have attempted to take a further step and diagnose effects of individual stressors in multiple-stressor environments, based on *a priori* predictions regarding the effects of

individual stressors on specific traits. Such studies have encountered problems when individual traits are responding simultaneously to multiple stressors (e.g. Doledec & Statzner 2008), or when the cause-effect relationship between a trait and a stressor is indirect and can be weakened by other environmental factors (e.g. Doledec et al. 2006). However, field data have matched predictions in cases where there is a strong, direct cause-effect relationship between a stressor and a trait (or traits; e.g. Doledec et al. 2006) or where a large number of consistently defined and described traits and trait categories have been used (e.g. Bonada et al. 2007). Schuwirth et al. (2015) further note that the ability of traits to diagnose individual stressors is weakened when traits indicating different stressors are correlated across taxa, or in studies where environmental stressors are strongly correlated. However, if all these potential problems are accounted for with appropriate selection of traits, sites and null hypotheses, reliable interpretations of trait responses can be achieved even in a multiple-stressor environment (Statzner & Beche 2010; Lange et al. 2014).

For a macroinvertebrate index to be useful for assessing human impacts at a national scale, it must be consistent across broad spatial scales and over time. Traits may be more consistent than taxonomic composition in these ways. For example, across the mainland United States, the percentage of aquatic macroinvertebrates classified as clinging taxa exhibited a consistent negative response to a gradient of increasing fine sediment, whereas the response of a taxonomic metric, EPT richness, varied significantly among geographic regions (Pollard & Yuan 2010). Therefore, a trait approach can potentially allow findings on stressor impacts to be transferred across geographic locations even where taxonomy differs due to biogeographic influences. Further, biological traits appear to be more stable among seasons than taxonomic composition (Beche et al. 2006). Thus, adopting a traits-based approach could potentially reduce biomonitoring sampling effort, as there is less variability to be accounted for via replication.

In theory, the traits-based approach may yield greater sensitivity than the taxonomic approach. This is because an assemblage's trait composition (sublethal changes and shifts in body size, or reproduction that occur without, or prior to, a loss of taxa) may show significant change before its taxonomic composition does (Culp et al. 2011). If so, a traits-based bioassessment may be able to detect changes in ecological health at mildly impacted sites better than a taxonomically-based assessment. In practice, traits have been more strongly related to differences human-related disturbances (such as land use) than a traditional taxonomic approach in some studies (Doledec et al. 1999, 2006) whereas in others (e.g. Lange et al. 2014) they have provided a similar or lower level of discrimination. The sensitivity of the traits-based approach is currently limited by the amount of traits information available (which also limits the taxonomic resolution at which it can be applied), so greater sensitivity could be anticipated as more information becomes available.

2.5.2. Method for deriving trait composition

The trait composition of a biological community is derived by combining its taxonomic composition with the trait 'profile' of each taxon present. The trait profile of a taxon can be represented by its affinity to different categories (or modalities) within each trait (see examples for categories and affinity scores in Tables 2 and 3, respectively). Affinities in the New Zealand traits database are coded using integers from 0 to 3, representing the strength of affinity for each category. For a variety of reasons taxa can have affinity scores > 0 for more than one category. For example, in terms of feeding, this can reflect that (1) a taxon changes feeding preferences during different stages of maturity, (2) a taxon shows different feeding behaviour in different environments, and/or (3) a taxon encompasses species that differ in their feeding preferences.

Table 2. Possible categories (modalities) within three example traits.

Trait	**Category**					
	1	**2**	**3**	**4**	**5**	**6**
Body size	≤ 5 mm	5–10 mm	10–20 mm	> 20–40 mm	> 40 mm	
Feeding	Shredder	Scraper	Deposit-feeder	Filter-feeder	Predator	Algal piercer
Attachment to substrate	Swimmer	Crawler	Burrower	Attached		

Table 3. Possible affinity scores of three taxa for the feeding trait based on categories defined in Table 2.

Taxon	**Categories**			
	Scraper	**Shredder**	**Predator**	**Filterer**
Taxon 1	3	1	0	0
Taxon 2	0	0	0	3
Taxon 3	3	0	2	0

The full array of affinity scores for all taxa across all trait categories is called a trait database. Trait composition of a community is typically described in terms of trait relative abundance. Trait relative abundance for a dataset was calculated as follows. First, a taxa-by-trait (trait modalities) matrix was prepared containing trait information, if available, for all taxa that occurred in the dataset. Trait data are not always available at the taxonomic level of identification. To simplify calculations, we converted all data to the taxonomic level required for the MCI (typically genus level) and then assigned the respective trait modality affinity scores if available. In several cases, trait

information was given at the species level while the taxa in our dataset typically were at genus level. Here the genus was assigned an average trait score across the existing scores for the species within that genus. The affinity scores in the taxa-by-trait matrix were standardised to sum to 1 for each trait. Secondly, a site-by-taxa matrix was prepared containing the relative abundances expressed as percentages of all taxa that occurred in the dataset. The relative abundances were log(x+1)-transformed to increase the relative contribution of naturally less-abundant taxa to the calculation of trait relative abundances. Addition of a constant value of 1 resulted in zero values for zero abundances and prevented undefined values after log transformation. Thirdly, matrix multiplication of the site-by-taxa and taxa-by-trait matrix result in a site-by-trait matrix. A fourth and final step is to divide the abundance of each trait modality by the sum of all modalities for that trait so that the modalities are expressed as relative abundances.

2.5.3. The New Zealand species traits database

A trait database for New Zealand taxa was first developed in 2004, with several updates until 2012 (Phillips & Reid 2012a, 2012b). The database is maintained by NIWA. However, some inaccuracies remained and a large number of knowledge gaps have since been filled by inference from international literature. As part of the current project, the New Zealand trait database was reviewed and updated using recent information and specialist knowledge of the major aquatic insect orders. Affinity scores for sixteen traits with between two and five modalities each (Table 4) were assigned to each taxon at the lowest level of identification available. The database now represents the best specialist knowledge available for New Zealand benthic macroinvertebrate fauna.

Table 4. List of species traits modalities of 16 traits assigned to New Zealand taxa.

Trait	Modality code	Modality description
Maximum potential size	SIZE1	≤ 5 mm
	SIZE2	> 5–10 mm
	SIZE3	> 10–20 mm
	SIZE4	> 20–40 mm
	SIZE5	> 40 mm
Maximum number of descendants per reproductive cycle	DESC1	≤ 100
	DESC2	> 100–1000
	DESC3	> 1000–3000
	DESC4	> 3000
Maximum number of reproductive cycles per year	SEMI	semivoltine
	UNIV	univoltine
	PLURIV	plurivoltine
Number of reproductive cycles per individual	CPI1	1
	CPI2	≥ 2
Life duration of adults	LDA1	≤ 1 day
	LDA2	> 1–10 days
	LDA3	> 10–30 days
	LDA4	> 30–365 days
	LDA5	> 365 days
Reproductive technique	SINGLE	single individual
	HERMA	hermaphrodism
	TWO	male and female
Oviposition site	SURFACE	water surface
	SUBMERGED	submerged
	TERRESTRIAL	terrestrial
	EGGENDO	eggs endophytic
Egg/egg mass	EGGFREE	free
	EGGCEMENT	cemented
	EGGPROTECTED	female bears eggs in/on body
Dissemination potential (all stages)	DISSLOW	low (10 m)
	DISSMEDIUM	medium (1 km)
	DISSHIGH	high (> 1 km)
Attachment to substrate of aquatic stages (excluding eggs)	SWIMMER	swimmers (water column)
	CRAWLER	crawlers (epibenthic)
	BURROWER	burrowers (infauna)
	ATTACHED	attached
Body flexibility	NOFLEX	none (< 10°)
	LOWFLEX	low (> 10-45°)
	HIGHFLEX	high (> 45°)

Table 4, continued

Trait	Modality code	Modality description
Body form	STREAMLINED	streamlined
	FLATTENED	flattened (dorso-ventral or lateral)
	CYLINDRICAL	cylindrical
	SPHERICAL	spherical
Feeding method	SHREDDER	shredders
	SCRAPER	scrapers
	DEPOSIT	deposit-feeders
	FILTERFEED	filter-feeders
	PREDATOR	predator
	ALGALP	algal piercer
Dietary preferences	SPECIALIST	strong (specialist)
	MODERATESPE	moderate
	GENERALIST	weak (generalist)
Respiration of aquatic stages (not including eggs)	TEGUMENT	tegument
	GILL	gills
	PLASTRON	plastron
	AERIAL	aerial
Aquatic stages	ADUANDLAR	adult, larva
	ADUORLAR	adult or larva
	LARANDPUP	larva, pupa

2.5.4. Trait responses to stressors

While several studies have examined trait responses to toxic contaminants (Liess & Von der Ohe 2005; Liess et al. 2008), only a few have examined responses to common stressors in agricultural environments. These have achieved various degrees of success in using traits to indicate agricultural stressors. In the western United States of America, Carlisle and Hawkins (2008) found traits of individual invertebrate taxa could be used to predict whether those taxa increased or decreased under altered land use, but were unable to predict catchment land use from the trait states present in invertebrate assemblages. In New Zealand, Doledec et al. (2006) showed that overall trait composition was correlated with specific agricultural stressors such as nutrient concentrations (dissolved inorganic nitrogen and dissolved reactive phosphorus) and deposited fine sediment. In addition, they showed that several specific traits responded to the overall gradient of agricultural intensity. They did not, however, report on the response of individual traits to individual stressors. In contrast, Lange et al. (2014) studied individual and combined effects of agricultural stressors on macroinvertebrate traits in a study of 43 stream sites along gradients of farming intensity (0–95% of the catchment in intensively managed grassland) and water

abstraction (0–92% streamflow reduction). Their findings indicated that several traits may be especially suitable for detecting effects of farming intensity because they all showed relationships along this gradient but not to the gradient in water abstraction. Similarly, other trait-based metrics such as the proportions of deposit feeders, scrapers and predators, look promising for indicating the effects of water abstraction because none responded to changes in farming intensity. Traits found to respond to specific stressors or gradients of agricultural intensity in previous studies are listed in Table 5. With regard to agricultural intensity and water abstraction, using traits as diagnostic tools requires describing and measuring more precisely the actual stressors that individual organisms experience in these altered waterways (Carlisle & Hawkins 2008).

Table 5. Traits identified as responding to agricultural stressors. References: 1. Pollard & Yuan 2010; 2. Richards et al. 1997; 3. Doledec et al. 2006; 4. Lange et al. 2014.

Stressor	Trait (or trait-based metric)	Predicted response to stressor increase	Trait modality in NZ trait database	Reference
Fine sediment	Clinging taxa relative richness	-ve	CRAWLER	1
	Merovoltine (≥ 3-year life cycle)	-ve	SEMI	2
	Multivoltine	-ve	PLURIV	2
	Large body size	-ve	SIZE4, SIZE5	2
	Scrapers	-ve	SCRAPER	2
	Clingers	-ve	CRAWLER	2
	Burrowers	+ve	BURROWER	2
	Lay unattached eggs at water surface	-ve	SURFACE	3
Intense agricultural land use (as DIN, DRP and %fines)	Univoltinism	-ve	UNIV	3
	Short adult life duration	-ve	LDA1, LDA2	3
	Male-female reproduction	-ve	TWO	3
	Free eggs	-ve	EGGFREE	3, 4
	Lay eggs at water surface	-ve	SURFACE	3, 4
	Low body flexibility	-ve	LOWFLEX	3, 4
	Filter feeders	+ve	FILTERFEED	3
	Plastron/aerial respiration	+ve	PLASTRON, AERIAL	3
	No or high dissemination	+ve	DISSLOW, DISSHIGH	3, 4
	≥ 2 reproductive cycles/individual	+ve	CPI2	3
	Hermaphroditism	+ve	HERMA	3
	Asexual reproduction	+ve	SINGLE	3
	Lay eggs below water surface	+ve	SUBMERGED	4
	Protect eggs on/in female body	+ve	EGGPROTECTED	3, 4
	Flexible, streamlined body shape	+ve	HIGHFLEX, STREAMLINED	3, 4
	Medium body size	-ve	SIZE2	4
	Attachment to substrate	-ve	ATTACHED	4
	Aquatic stages	-ve	ADUORLAR	4

Table 5, continued

Stressor	Trait (or trait-based metric)	Predicted response to stressor increase	Trait modality in NZ trait database	Reference
Water abstraction	Scrapers	-ve	SCRAPER	4
	Respiration of aquatic stages through gills	-ve	GILL	4
	Deposit feeders	+ve	DEPOSIT	4
	Predators	+ve	PREDATOR	4
	Moderate dietary preference	+ve	MODERATESPE	4
	Respiration of aquatic stages through tegument	+ve	TEGUMENT	4
	Life duration of adults (>365 d)	+ve	LDA5	4

Two cautions must be noted when using traits to identify specific stressors. First, the potential for traits to correlate may confound causal inference in biomonitoring applications (Schuwirth et al. 2015). As a result, it is possible that a limited number of possible trait combinations or 'syndromes' may exist (Poff et al. 2006; Horrigan & Baird 2008), and correlations among trait-based indices may occur partly due to correlations among traits across taxa. Such correlations need to be quantified in different environments so their effect can be accounted for (e.g. Schuwirth et al. 2015) and the relative sensitivity of individual and correlated traits to different stressors needs to be evaluated more fully. In some cases correlated traits may be redundant (Poff et al. 2006).

The second caution applies to multiple-stressor environments. Where a mechanistic linkage between traits and environmental stressors can be established, the relationship can become confounded if particular traits respond to multiple features of the environment (Statzner & Beche 2010), or if multiple environmental stressors are correlated among sites (Yuan 2007; Schuwirth et al. 2015). In one study in a multiple-stressor environment (Horrigan & Baird 2008), some trait modalities were influenced exclusively by changes in flow conditions and were not responsive to thermal and oxygen stress, whereas other traits were simultaneously responsive to the multiple stressors, and consequently had reduced diagnostic power (Yuan 2007). Consequently, Culp et al. (2011) recommended more research to identify stressor-specific traits (trait suites) based on the understanding of the causal relationship between trait occurrence and stressor level, and Yuan (2007) recommended using multiple traits rather than single trait to infer environmental stressors. Lange et al. (2014) provided a regional example of such an analysis. Future tasks in this project will test whether individual traits (or trait-based metrics) respond to one or more stressors within a multiple stressor environment at a national scale (Section 5).

3. DEVELOPMENT TOWARDS NEW STRESSOR-SPECIFIC METRICS

3.1. Overview

The aim of this task was to develop macroinvertebrate indices relevant to four major pressures in New Zealand rivers: sedimentation, eutrophication, habitat modification and change in flow due to abstraction pressures. This section describes several steps that were undertaken working towards this aim, specifically:

1. a systematic review of the literature characterising and, where possible, quantifying taxon-specific and trait-specific responses to stressors (Section 3.2) focussing on in-stream sediment
2. an estimation of tolerance values for taxa in response to specific stressors based on information from the literature and best professional judgement, including deposited sediment, nutrients via a periphyton causal pathway, temperature, oxygen and metals (Section 3.3).
3. the calculation of 20 new sediment-specific and nutrient-specific stressor metrics using tolerance values derived from statistical analysis of the collated research dataset described in Section 2.1.2 (Section 3.4).

3.2. Introduction

Various stressor-specific macroinvertebrate indices have been developed overseas as a tool for measuring human impacts at the reach scale and determining ecologically relevant site-specific targets for, for example, sedimentation (Bryce et al. 2010; Extence et al. 2013), nutrient enrichment (Smith et al. 2007; Haase & Nolte 2008), pesticides (Liess & Von der Ohe 2005), low flows and water abstraction (Extence et al. 1999). These indices have been developed using a variety of approaches from expert judgement to complex statistical modelling.

Most commonly, tolerance scores are assigned to species based on expert judgement, or informed by empirical data, and these tolerance scores are used to develop stressor-specific metrics. For example, the Proportion of Sediment-sensitive Invertebrates (PSI) is based on fine sediment sensitivity ratings of species and families of British benthic macroinvertebrates (Extence et al. 2013). In contrast, tolerance values in the Deposited Sediment Biotic Index (DSBI) were assigned based on the deposited-sediment level at which the taxon reached 50% cumulative abundance across a Missouri, USA data set (Zweig & Rabeni 2001). Among the examples given above, the pesticide-specific metric SPEAR (SPEcies At Risk) is an exception as it is based on biological traits responsive to the effects of pesticides rather than taxonomy.

Furthermore, Threshold Indicator Taxa Analysis or 'TITAN' (Baker & King 2010) was developed to identify species sensitivity or tolerance to increasing stress and to identify community thresholds. A weakness of this approach is that models only take into account the effects of single stressors which could bias threshold values if the dataset contains multiple stressor as well as environmental gradients. This approach furthermore has been criticised from a statistical standpoint (Cuffney & Qian 2013). More specifically, TITAN accurately and consistently identified thresholds in a dataset that simulated responses to a disturbance gradient according to the step-function model but failed to do so in models characterised by abrupt changes in response slopes or response direction (Cuffney & Qian 2013). Furthermore, threshold identification with TITAN was very sensitive to the distribution of 0 values (Cuffney & Qian 2013). Finally, the proposed tests of statistical significance led to inflated estimates of statistical signficance and underestimates of the confidence intervals of the identified thresholds (Cuffney & Qian 2013).

Gradient forest is another approach that has been developed to identify community thresholds based on models for multiple taxa (Ellis et al. 2012) and recently applied to a New Zealand dataset (Wagenhoff et al. 2017b). These taxon models take into account the effects of multiple stressors or environmental drivers, hence, render the gradient forest analysis as potentially useful for assignment of tolerance values and development of stressor-specific indices which can be used to inform resource limits and management options. Limitations of these models to tease apart the effects of multiple drivers (when drivers are heavily correlated in the dataset) are discussed below (Section 3.5.4).

3.3. Systematic review of the literature

3.3.1. Introduction

Systematic reviews are in contrast to narrative reviews as they treat relevant literature as data (Khan et al. 2003), and employ statistical analysis to succinctly analyse and summarise a large body of literature, testing the level of support for hypotheses across numerous studies (Webb et al. 2015). A systematic synthesis improves the defence and transparency of decision making and may help increase scientific input into the setting of resource limits and freshwater targets/objectives (Webb et al. 2013). This would not only fulfil legal requirements to create 'evidence-based' environmental management, but could also in turn improve environmental outcomes. However, systematic reviews can be highly resource intensive, which may have limited their use to date.

3.3.2. Methods

We conducted a systematic review of the literature pertaining to sediment effects on benthic macroinvertebrates using Eco Evidence software (Webb et al. 2015). The review was conducted as part of the parallel MfE-funded project on sediment attributes (Depree et al. 2017). Methodological details of the Eco Evidence systematic literature review are provided in Appendix 2. References for the 65 studies interrogated using the Eco Evidence approach are also given in Appendix 2. A brief outline of the results and key findings of the Eco Evidence approach is provided in this section, focussing on the findings that support the development of a sediment-specific macroinvertebrate metric.

3.3.3. Results

Overall, 655 cause-effect hypotheses were tested. This large number reflects all the possible species, traits and metrics combined with all the possible measures of in-stream sediment. Most hypotheses had insufficient evidence (weighted data) to test the cause-effect relationship. However 111 hypotheses had sufficient evidence to support the hypothesis, support an alternate hypothesis, or inconsistent evidence to support the main or alternate hypothesis (Table 6).

3.3.4. Key findings

The Eco Evidence systematic review confirmed 25 hypotheses (original or alternate) of the effect of sediment on benthic macroinvertebrates. In response to a general increase in deposited fine sediment, 14 hypotheses were supported by the meta-analysis of the literary data including a decrease in 8 taxa, 3 species traits and 3 community metrics. In particular, EPT metrics (i.e. EPT density, %EPT abundance) were a good indicator of deposited fine sediment effects. There was also significant evidence of the effect of deposited fine sediment on the MCI metric. Eleven alternate hypotheses were supported by the analysis including an increase in 2 taxa, 1 trait and 1 metric, and a decrease in a further 4 taxa and 3 traits (Table 6).

There was little consistency among responses when comparing patch-scale and reach-scale measures of deposited fine sediment, other than for decreases in EPT richness and abundance (Table 6). This brings into question whether causality versus correlation was the main driver of relationships observed in many studies. There was no overlap between deposited sediment and suspended sediment in supported hypotheses. An increase in suspended sediment causing a decrease in macroinvertebrate abundance was the only causal relationship supported by the literature for suspended sediment.

Table 6. Results of the Eco Evidence review showing cause-effect hypotheses that contained sufficient evidence from the literature to reach an outcome of support for the hypothesis, support for an alternate hypothesis or inconsistent evidence. Arrows indicate the direction of response. * = taxa not found in New Zealand.

Treatment Metric	Support	Alternate	Inconsistent
↑**Deposited fine sediment**	↓%EPT abundance ↓clinger ↓*Deleatidium* ↓*Ecdyonurus** ↓Elmidae ↓Ephemeroptera ↓EPT density ↓*Leuctra** ↓low body flexibility ↓MCI ↓Orthocladiinae ↓*Paraleptophlebia** ↓Plecoptera ↓surface egg laying	↑Baetidae* ↑macroinvertebrate biomass ↑*Potamopyrgus antipodarum* ↑respires using gills ↓%crawlers ↓Cladocera ↓Copepoda ↓*Oxyethira* ↓scraper ↓shredder ↓Tanypodinae	↑burrower ↑*Hexatoma** ↑macroinvertebrate density ↑Nematoda ↓%EPT ↓Chironomidae ↓EPT abundance ↓EPT richness ↓filter-feeder ↓*Glossosoma** ↓*Hesperoperla pacifica** ↓macroinvertebrate abundance ↓macroinvertebrate diversity ↓macroinvertebrate richness ↓Oligochaeta
↑**% cover**	↑burrower ↓%EPT abundance ↓clinger ↓*Deleatidium* ↓Ephemeroptera ↓EPT density ↓low body flexibility ↓MCI ↓*Paraleptophlebia** ↓Plecoptera ↓surface egg laying	↑Baetidae* ↑macroinvertebrate biomass ↑*Potamopyrgus antipodarum* ↓%crawlers ↓Cladocera ↓Copepoda ↓Oligochaeta ↓scrapers ↓shredders ↓Tanypodinae	↑*Hexatoma** ↑macroinvertebrate density ↑Nematoda ↓%EPT ↓Chironomidae ↓EPT abundance ↓EPT richness ↓*Glossosoma** ↓macroinvertebrate abundance ↓macroinvertebrate diversity ↓macroinvertebrate richness ↓*Neophylax**
↑**% cover (measured at the patch scale)**	↑burrower ↑nematoda ↓%EPT ↓*Deleatidium* ↓Ephemeroptera ↓EPT abundance ↓EPT density ↓*Paraleptophlebia** ↓Plecoptera	↑Baetidae* ↑*Potamopyrgus antipodarum* ↓Cladocera ↓Copepoda ↓Tanypodinae	↑macroinvertebrate density ↑Oligochaeta ↓Chironomidae ↓EPT richness ↓macroinvertebrate abundance ↓macroinvertebrate diversity ↓macroinvertebrate richness ↓*Neophylax** ↓scrapers ↓shredders
↑**% cover (measured at the reach scale)**	↓%EPT abundance ↓EPT density ↓EPT richness ↑macroinvertebrate density	↓Chironomidae ↓macroinvertebrate diversity ↓Oligochaeta ↓shredder	↓macroinvertebrate abundance ↓macroinvertebrate biomass ↓macroinvertebrate richness
↑**Suspended sediment**	↓macroinvertebrate abundance		↓macroinvertebrate richness ↓EPT richness

Overall, several taxa and traits showed a consistent response to fine sediment and this provides further support for the development of a sediment-specific metric. However, results showed overwhelmingly that the majority of hypotheses had insufficient evidence illustrating that there remain significant knowledge gaps and any decisive statements from narrative reviews should be read with caution.

3.4. Identification of tolerance values based on best professional judgement

The project team undertook a group exercise (during Workshop 1) in assigning tolerance scores for macroinvertebrate taxa based on their expert knowledge of the expected response shape to deposited sediment. Taxa were graded into four categories based on team consensus (Table 7). After the workshop, seven experienced team members were tasked to assign their own taxa sensitivity scores again for deposited sediment, and also for nutrients via a periphyton causal pathway, temperature, oxygen and metals. A summary of these scores are presented in Appendix 3, Table A3.1. Comparison of these taxon scores suggested that stressor-specific metric development for the major stressors nutrients and deposited sediment is unlikely to be successful. Many taxa that had high sensitivity scores for nutrients also had high scores for sediment and similarly for low sensitivity scores. We think that it was likely hard for experts to tease apart the effects of these two stressors. Hence, we decided that our primary approach to metric development would be the calculation of sensitivity scores from available paired stressor-taxa datasets (Section 3.5) and the tolerance scores developed through expert knowledge from seven team members were not used to derive metrics in this study. Scores provided in Appendix 3 could be used in future studies.

Table 7. Tolerance scores assigned to taxa/traits based on expert opinion.

Score	Description	Expected response shape
A	highly sensitive	decrease
B	moderately sensitive	subsidy-stress or slight decrease
C	moderately insensitive/tolerant	no response
D	highly insensitive/tolerant/favoured	increase

3.5. Stressor-specific metric development

3.5.1. Introduction

Stressor-specific metrics are useful tools for managing ecosystem health because they are diagnostic tools that assist in identifying the main cause(s) of stream ecosystem health degradation at specific sites and/or at a regional/national level, and potentially aid in identifying the limiting factor(s). For rehabilitation projects or regional policy development, such knowledge would help in deciding what stressor(s) to address first in order to reach positive ecological outcomes. Once causes and limiting factors have been identified, stressor-specific metrics could be used to track restoration success or effectiveness of regional policies over time with respect to management of a specific stressor.

Stressor-specific metrics are expected to show a strong response and good relationship with each respective stressor. Hence, models of the response shape of these metrics across stressor gradients in a spatial dataset could be used to identify thresholds that help with definition of instream objectives (targets) for manageable stressors, e.g. development of an attribute within the NPS-FM.

Overview to metric development and rationale for using a gradient forest analysis

There are three main steps to metric development: (1) identification of sensitive or tolerant taxa and/or assignment of tolerance values to taxa, (2) combining the tolerance values into a metric score, and (3) validating the metric, i.e. testing whether the metric responds to the stressor that it was developed for and, for stressor-specific metrics, whether they are able to discriminate between the different stressors. During this project, we tested the proof of concept of stressor-specific metrics. Further work with respect to refining tolerance values and metric calculation as well as metric validation will likely be needed in order to develop metrics that can be fully implemented into management and policy.

Gradient forest (GF) analysis was chosen to inform Step 1 of metric development, i.e. assignment of tolerance values to macroinvertebrate taxa. Gradient forest is a relatively new approach developed for calculating the importance of gradients and identifying assemblage thresholds using regression tree-based random forest (RF) models for individual taxa (Ellis et al. 2012). For example, this approach has recently been used to identify congruence in stream assemblage thresholds among macroinvertebrate, periphyton and bacterial assemblages in response to nutrient and sediment gradients (Wagenhoff et al. 2017b). Here, however, we adopted the analytical approach to inform tolerance value development. To our knowledge, gradient forest has never been applied for this purpose. We anticipate that this approach is superior to single-stressor models in describing taxon responses to specific stressor gradients. However, this approach will be new to the wider scientific community.

The strength of model approaches such as RF is the ability to model complex, nonlinear response shapes and the inclusion of multiple predictors that strengthen the evidence of cause and effect between stressors and taxon responses. Furthermore, potential complex interactions among stressors and also natural environmental variables are automatically handled (Cutler et al. 2007). These characteristics are promising for disentangling the effects of multiple stressors (here sediment and nutrients) and natural environment gradients typically prevalent in broad spatial datasets. However, the ability to disentangle the effects of multiple drivers using a spatial dataset depends on correlations between drivers within that dataset. While RF models are capable of dealing with correlated predictors pretty well, it is unknown at what degree of correlation the interpretation of the results, i.e. assignment of tolerance values, becomes less reliable. Validation of the metrics using independent spatial datasets that also contain minimum correlation between these stressors or using experimental datasets would be best to test the validity of stressor-specific tolerance values and hence metrics.

Description of the gradient forest analysis

The GF method is described in detail by Ellis et al. (2012). It is performed in the freely-available statistical programme R (R Core Team 2016, here in R version 3.3.2) using two R packages provided by the same authors. A shorter description of the computational method of the GF analysis available in Wagenhoff et al. (2017b) is given below.

First, R package 'extendedForest' builds RF models for each taxon consisting of 500 regression trees. Each regression tree is fitted to a bootstrap sample of the observations and partitioning of the data is performed by the best split (minimising error variance) tested on a random subsample of the predictors (Cutler et al. 2007). Each split is associated with an importance value reflecting the degree of change in abundance. The RF model predictions are averages of the predictions of each tree. The goodness-of-fit measure R^2 (pseudo R^2) is the proportion of the variance explained by the RF model and derived through cross-validation, i.e., estimated from the error variance of the out-of-bag sample process (Ellis et al. 2012; Wagenhoff et al. 2017b). The package extendedForest calculates an improved, more robust measure of predictor importance within each RF model by taking into account correlation between predictors. We used a default correlation threshold of 0.5.

Secondly, the R package 'gradientForest' uses information only of those taxon-specific RF models with $R^2 > 0$. For tolerance value assignment, we used RF forest model output as well as the GF output that computes species turnover functions (i.e., for each individual taxon). These species turnover functions are computed using information on the predictor splits and associated importance within a tree by accumulating the split importance values across each environmental gradient. The GF output also provides the importance of each predictor for overall compositional turnover, which is calculated by taking a weighted average of the taxon-specific

conditional predictor importance using the R^2 values of the RF models (Ellis et al. 2012).

3.5.2. Gradient forest analysis method

Research data selection

We selected a subset of the research dataset containing 1,861 samples collected from 973 sites (Section 2.2.2) using a range of criteria. First, we excluded all data from mesocosm experiments with the rationale that there are spatial and temporal limitations that may make these samples not entirely comparable with samples taken in real streams. For the same reasons, we also excluded samples from field experiments that were taken after addition of sediment and/or nutrients. Samples from these field experiments taken before treatment or taken at control sites, however, were selected for GF analysis.

Secondly, for three studies where samples had been taken at the same sites for multiple years to study the effects of different catchment land-use scenarios, only a subset of these samples was selected for analysis to avoid bias towards these sites. We selected up to five samples from each of those sites to balance the inclusion of as many data points as possible and reduction of bias. Samples from each site were selected to span a gradient in sediment and/or nutrients which was likely due to changes over time in catchment land use. In two studies where samples were taken in all seasons or in summer and winter, selections were made only from summer or autumn samples. In one study, where there were three sampling occasions in a single year, we selected the one when sediment was estimated in two different ways.

Finally, GF analysis requires that there are no missing values in the dataset. Hence, the number of data points used in the analysis depends on which stressor attributes will be used as predictors. Sites with only sediment data but not nutrients or periphyton or vice versa did not make it into the gradient forest analysis.

Stressor attributes

For sediment, we selected two attributes that are commonly used to describe sedimentation (1) percentage cover of fine sediment on the streambed, and (2) suspendable inorganic sediment (SIS) mass on the streambed. Percentage of sediment cover was assessed mainly instream. In order to maximise the number of samples from the large dataset to go into the GF analysis, we also accepted an estimate of ‘percentage’ sediment cover calculated from a Wolman pebble count of typically 100 particles (SAM3, Clapcott et al. 2011).

We selected dissolved inorganic nitrogen (DIN) as the most commonly measured nitrogen attribute, but also accepted the sum of NO_3-N and NH_4-N as well as an assessment of NO_3-N only because NO_2-N and NH_4-N typically make up a small

portion of DIN. We selected dissolved reactive phosphorus (DRP) and chlorophyll-*a* as the most commonly measured phosphorus and periphyton attribute, respectively.

Environmental covariates

Several catchment- and reach-scale environmental descriptors were chosen as they have previously been shown to influence benthic macroinvertebrate communities (e.g. Clapcott et al. 2012; Wagenhoff et al. 2017). These included variables representing flow, temperature, geology, catchment morphology, position in the stream network, substrate size, shading and elevation (Table 8). The rational for including environmental covariates was to be able to account for their effects and hence isolate stressor effects from confounding factors. The limitations of the ability to disentangle effects of individual stressors as well as environmental factors has been discussed in Section 3.4.1.

Table 8. Set of 19 predictor variables used in RF models along with their data source and description. Measured data come from the large research data set compiled during this project (Section 2.1.2). Three flow statistics (Booker 2013; Booker & Woods 2014) were downloaded on 23 August 2016 from the MfE website (https://data.mfe.govt.nz/table/2536-natural-river-flow-statistics-predicted-for-all-river-reaches/); REC = River Environment Classification database (Snelder & Biggs 2002), FENZ = Freshwater Ecosystems New Zealand database (Leathwick et al. 2011).

Predictor	Source	Description
sed_SIS	Measured	Suspendable inorganic sediment (g/m^2)
sed_cover	measured	% sediment cover visually estimated instream
peri_chl.a	measured	Benthic chlorophyll-*a* from rock scrapings
nut_DIN	measured	Dissolved inorganic nitrogen
nut_DRP	measured	Dissolved reactive phosphorus
ORDER	REC	Stream order
ELEVATION	REC	Altitude of the stream segment
SegJanAirT	FENZ	Summer air temperature for a segment
SegMinTNor	FENZ	Seasonal air temperature range for a segment
SegRipShade	FENZ	Riparian shade for a segment
USCalcium	FENZ	Average calcium concentration of underlying rocks
USPhosphorus	FENZ	Average phosphorus concentration of underlying rocks
USHardness	FENZ	Average hardness of underlying rocks
LocSed	FENZ	Weighted average of proportional cover of bed substrate of different size categories
USSlope	FENZ	Average slope in the catchment
DSDist	FENZ	Distance to coast
SegFlowStability	FENZ	Ratio of mean annual low flow/ mean annual mean flow
SpecMeanF	MfE website	Specific mean flow (= mean flow / catchment area)
SpecMALF	MfE website	Specific mean annual low flow (= mean flow / catchment area)
FRE3	MfE website	Annual frequency of flood events > 3x median annual flow

Gradient forest model settings

Two GF analyses were run, one using SIS as a sediment predictor and the other using % sediment cover. Investigation of taxon responses to both sediment attributes had the advantages that (1) overall, more data points made it into the analysis and that should provide a better coverage of stressor and environmental gradients in the multidimensional space, hence improving predictive accuracy; and (2) SIS and % sediment cover may provide complementary information on the impacts of sedimentation. Random forest analysis requires that there are no missing values in the dataset. This requirement substantially reduced sample size compared to the dataset containing the selected sites (N = 812) from the large dataset. For the GF analysis using SIS or % sediment cover as the sediment predictor, sample size was reduced to 161 and 306, respectively.

Macroinvertebrate data were log-transformed to approximate normal error distributions using the formula ln(x + min(x > 0)) where x is the relative abundance expressed as a proportion. Analysis was restricted to taxa that had at least 10 non-zero values (occurrences), i.e. were found in at least 10 samples. Sixty-six of the total of 116 taxa met this cut off for the 'SIS' macroinvertebrate dataset, and 78 of 128 taxa met the cut off for 'sedcover' macroinvertebrate dataset. The GF approach first calculates RF models for each of these taxa. For all taxa with an RF model that had an $R^2 > 0$, i.e. for which our 19 predictors had at least some explanatory power, species turnover functions were calculated.

Identification of 'decreasers' and 'increasers' in response to a stressor gradient

The assumption is that taxa decreasing across the stressor gradient ('decreasers') are sensitive taxa while taxa increasing across the stressor gradient ('increasers') are taxa tolerant to the stressor, or even favour the stressor. Species turnover functions do not depict whether a taxon decreased or increased across the stressor gradient. Hence, sensitive and tolerant taxa were identified using both the modelled response shapes as well as expert opinion, as follows. First, the partial dependence plots of the RF models (with overlay of data points) were visually investigated, and taxa with response shapes that are either predominantly negative or positive were named decreasers or increasers, respectively. Taxa that showed neither of these patterns were classified as 'unclear', which may be a result of taxa being unresponsive to the stressor gradient or due to insufficient data. It was not obvious which reason was prevalent within our data. Random forest models are good predictive models but because they are not parametric models, parameter estimates (i.e. response shape) cannot be tested for statistical significance. Nevertheless, a clear positive or negative response shape provides some evidence for taxa increasing or decreasing to a stressor gradient, respectively. On the other hand, the decision of whether small, directional variations in the response shape are meaningful or due to unusual data points are somewhat subjective. Hence, as a second step expert opinion was provided, mainly by Jon Harding (University of Canterbury), as to whether the

assignment to these three categories simply based on model output also makes ecological sense.

Identification of thresholds

A threshold in this context was defined as the stressor value at which there is an abrupt change in the relative abundance of a taxon compared to other points across the stressor gradient. Thresholds were calculated from the species turnover functions across % sediment cover, SIS and chlorophyll-*a*. Note that these were derived from two different models, one for the 'SIS' and one for the 'sedcover' macroinvertebrate dataset. Thresholds for each taxon were identified at the value in a stressor variable where the species turnover function reaches 25% of the maximum cumulative importance. The 25% value was selected by the team based on visual assessment of where the most abrupt change in relative abundance generally occurred by comparing partial dependence plots with species turnover functions. Note that this step in the analysis provided a ranking of the sensitive as well as tolerant taxa, respectively, as a basis for assigning tolerance values. It is not to be confused with setting management objectives such as one that allows management up to a point where abrupt changes in taxon relative abundances occur. In a pilot analysis we had used the 50% value at which we identified that thresholds and it appeared that the ranking was not much influenced.

Tolerance value assignment

Tolerance values were assigned to further discriminate between different degrees of sensitivity or tolerance. Tolerance values for decreasers were assigned to a range from 10 to 6 (i.e. 10, 9, 8, 7 or 6) with values of 10 being assigned to the most sensitive taxa, i.e. those with the lowest thresholds. Tolerance values for increasers were assigned to a range from 1 to 5 (i.e. 1, 2, 3, 4 or 5) with values of 1 being assigned to the most tolerant taxa, i.e. those with the lowest thresholds. We investigated assignment of tolerance values based on the raw stressor gradient as well as the natural log-scaled stressor gradient because taxa often respond in a log-linear fashion to nutrient concentrations.

A 1-10 scale was chosen for tolerance assignment as managers in New Zealand are familiar with this scale, which has been used for tolerance value assignment for the most commonly-used stream health metric, the MCI. However, the analytical approaches to tolerance value assignment between the MCI and our approach fundamentally differ in that our approach uses stressor-response shapes whereas the approach adopted for the MCI is simply based on association of a taxon with an *a priori*-selected pollution gradient. Furthermore, we have not assigned tolerance values to the 'unclear' group of taxa which potentially could be assigned an intermediate tolerance value (e.g. 5).

Metric calculation and validation

We calculated the following assemblage-level metrics for each stressor using the tolerance values (Tables 8 and 9) and the formula for the MCI-type metrics provided in Section 2.3.1.

- Number of sensitive taxa (decreasers)
- % sensitive taxa richness (decreaser taxa / richness * 100)
- % sensitive taxa abundance (decreaser abundance / total abundance * 100)
- Number of tolerant taxa (increasers)
- % tolerant taxa richness (increaser taxa / richness * 100)
- % sensitive taxa abundance (increaser abundance / total abundance * 100)
- 'MCI' (raw-scale tolerance value assignment)
- 'MCI' (log-scale tolerance value assignment)
- 'QMCI' (raw-scale tolerance value assignment)
- 'QMCI' (log-scale tolerance value assignment).

Preliminary validation was performed by plotting the metrics against sediment and nutrient stressor gradients, and then performing a simple linear regression using the same dataset as used for metric development. Note that this analysis does not replace true validation which should be based on independent data.

3.5.3. Results

Gradient forest output

For the 'SIS' GF analysis, 52 out of the 66 taxa had RF models with $R^2 > 0$ with R^2 values ranging from 0.01 to 0.98 (for *Zephlebia*). For the 'sedcover' analysis, 65 out of 78 taxa had RF models with $R^2 > 0$ with R^2 values ranging from 0.01 to 0.88 (for *Zephlebia*). These are the taxa for which the GF approach calculates the overall predictor importance across all taxa (Figure 3) and the species turnover functions for all taxa which can be presented in species cumulative plots (Figure 4).

Predictor sed_SIS_gm2 ranked third in importance, while chlorophyll-*a* ranked 11th (Figure 3, left panel). In the second GF analysis for sedcover, predictor sed_cover_final ranked fourth while chlorophyll-*a* ranked 13 (Figure 3, right panel). We only interpreted the chlorophyll-*a* outcome from the 'sedcover' GF analysis. In future metric development, it would be useful to see if the relationships of the taxa with chlorophyll-*a* are the same when different sediment measures are used.

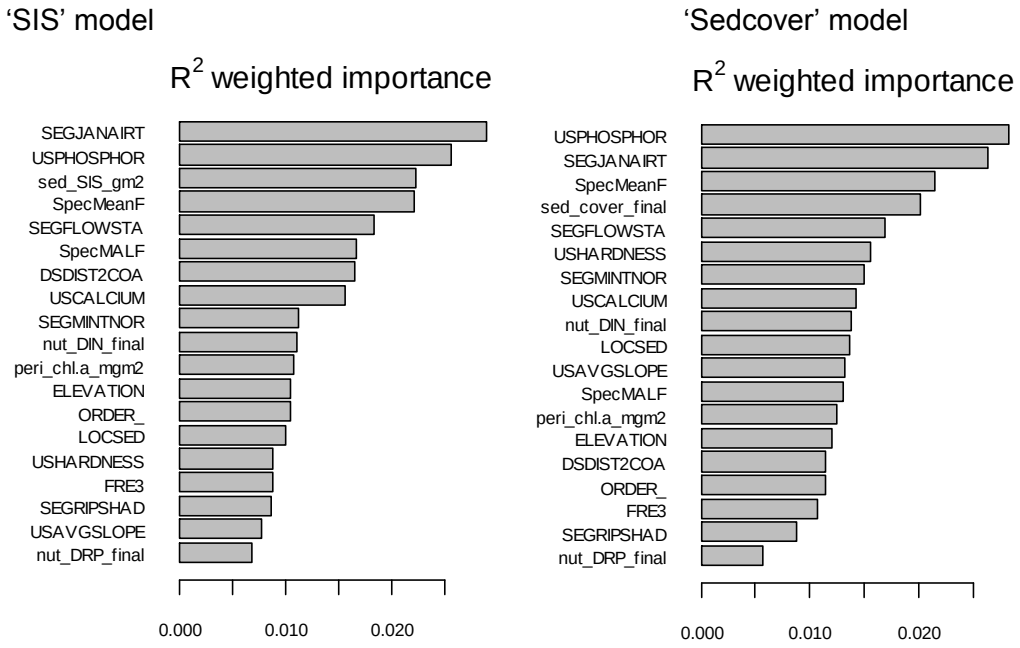


Figure 3. Overall predictor importance (in R^2 units) for species distribution, calculated by gradient forest (GF) analysis, allowing assessment of the relative importance of the environmental predictors for the macroinvertebrate assemblage. The left panel shows the results of the 'SIS' GF analysis while the right panel shows the results of the 'sedcover' analysis.

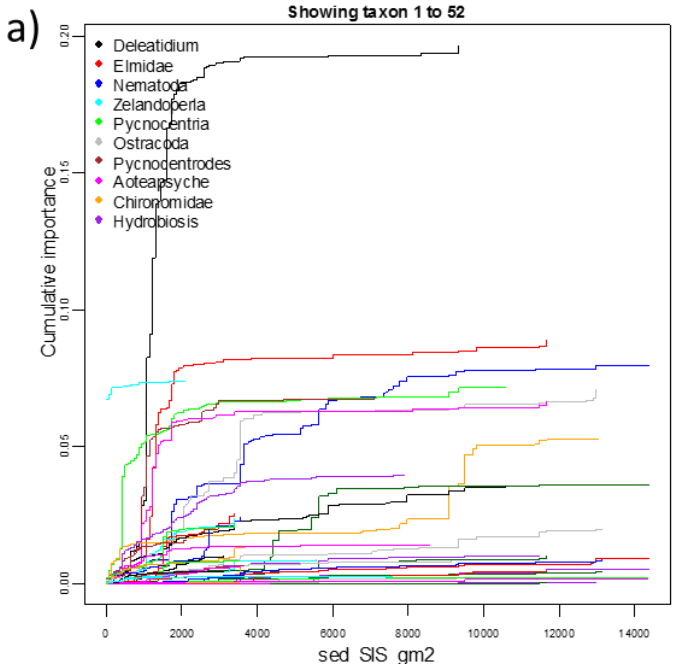
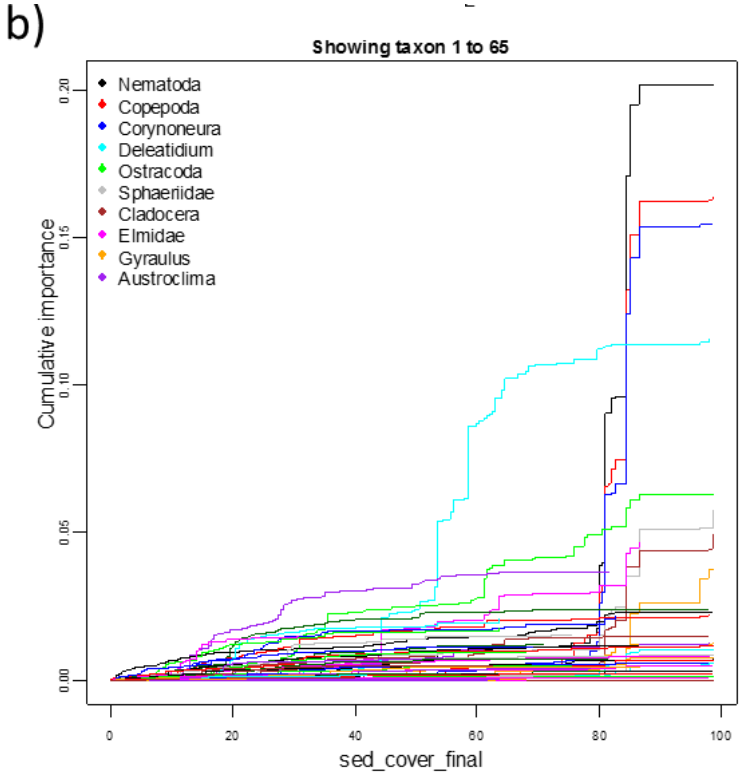


Figure 4. Species turnover functions for all taxa that had a RF model with $R^2 > 0$. The results for the SIS gradient are from the 'SIS' GF analysis, while the results for the sediment cover and chlorophyll-*a* gradient are from the 'sedcover' GF analysis. SIS = suspendable inorganic sediment (in g/m^2), sed_cover_final = % sediment cover, peri_chl.a_mgm2 = chlorophyll-*a* (in mg/m^2). The legend only lists the 10 most important taxa ordered by their maximum cumulative importance (in R^2 units).

b)



c)

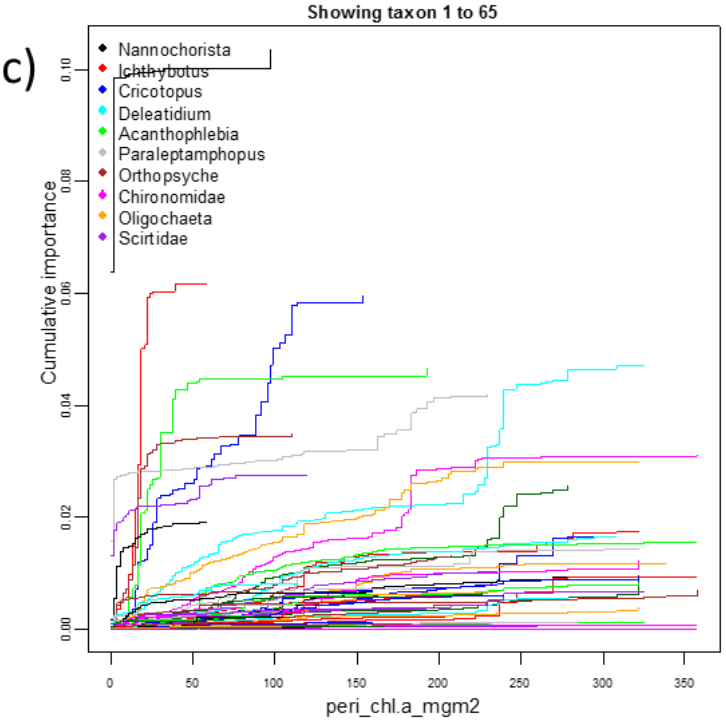


Figure 4, continued.

Note that the following results make use of the GF output but that the developers of the GF approach did not intend use of this approach for this purpose. We adopted the GF approach, for the first time, as a novel means of determining sensitive and tolerant taxa and further assigning tolerance values to these taxa. A validation procedure using an independent dataset would test the robustness of our approach.

Decreasers and increasers, thresholds and tolerance values

For the sediment-specific metric, the SIS and sediment cover gradients could be used to define response groups for each taxon (decreasers, increasers, unclear) and assign tolerance values. According to expert opinion, overall the response shapes were more correlated with conventional wisdom for the sediment cover gradient compared to the SIS gradient. Hence, the response groups were assigned to taxa based on the sediment cover gradient, however the response shapes to SIS were used to confirm the assignment if there was disagreement between the model-based and expert assignment. Consequently, the response shapes to chlorophyll-*a* were used from the 'sedcover' model.

For the sediment gradient, we identified 25 decreasers and 12 increasers while 28 taxa were classified as 'unclear' (Table 9). Sediment thresholds ranged from 5 to 82% for decreasers and from 12 to 82% for increasers. Assignment of tolerance values based on the raw scale of the sediment gradient resulted in the majority of taxa having the maximum possible score of 10 (20 out of 37) and few taxa with other scores (Table 9). The assignment based on the log-scale resulted in a somewhat more even distribution across the values of 1–10. In order for the stressor-specific metrics to have diagnostic power, the taxon list and/or response direction or tolerance values should differ between the two stressors. We identified indicator taxa (decreasers or increasers) for sedimentation that were not among the indicator taxa for eutrophication and vice versa (Tables 9 and 10). We also identified taxa that responded in the opposite direction for these two stressors.

Table 9. List of taxa along with model R^2, number of occurrences in the dataset, maximum cumulative importance, % sediment cover threshold defined at the point where 25% of the maximum cumulative importance is reached, and tolerance values assigned based on the raw-scale and log-scale sediment gradient. Note that the models of seven taxa were selected by the R package as having an $R^2 > 0$, however their R^2 values were undistinguishable from zero. * indicates taxa that were decreasers or increasers across the sediment but not across the periphyton gradient, ** indicates taxa that responded to sediment in the opposite way compared to periphyton. Taxa in grey were classified as unclear.

Taxon	No. of occurrences	Model R^2	Response group	Maximum cumulative importance	% Sediment cover threshold at 25% of max. cum. imp.	Tolerance value (raw-scale gradient)	Tolerance value (log-scale gradient)
Beraeoptera	22	0.09	decreaser	0.00001	5	10	10
Zelandoperla	51	0.19	decreaser	0.01151	5	10	10
*Costachorema**	33	0.25	decreaser	0.00490	8	10	10
*Stenoperla**	45	0.18	decreaser	0.00008	8	10	10
Chironomidae*	165	0.43	decreaser	0.02313	8.5	10	10
Helicopsyche	28	0.14	decreaser	0.00114	9	10	9
*Maoridiamesa***	41	0.37	decreaser	0.00689	9.5	10	9
*Archichauliodes**	124	0.35	decreaser	0.01219	10	10	9
Ameletopsis	10	0.33	decreaser	0.01464	11	10	9
Hydraenidae	34	0.28	decreaser	0.00054	11	10	9
Austroperla	37	0.17	decreaser	0.00470	11.5	10	9
Pycnocentria	127	0.12	decreaser	0.01168	11.5	10	9
*Nesameletus**	56	0.35	decreaser	0.00806	12	10	9
*Psilochorema**	140	0.19	decreaser	0.00805	12.5	10	9
*Acanthophlebia**	17	0.74	decreaser	0.00204	13.5	10	9
Orthopsyche	25	0.50	decreaser	0.00697	13.5	10	9
Ptilodactylidae*	11	0.57	decreaser	0.00225	14.5	10	9
Pycnocentrodes	169	0.14	decreaser	0.02213	14.5	10	9
*Latia**	18	0.03	decreaser	0.00041	15.5	10	8
*Olinga***	119	0.16	decreaser	0.01514	17	10	8
*Hydrobiosis**	197	0.18	decreaser	0.01262	23.5	9	8
*Aoteapsyche**	206	0.29	decreaser	0.01480	31.5	9	7
Elmidae*	251	0.40	decreaser	0.04678	44	8	7
Deleatidium	263	0.49	decreaser	0.11550	52.5	7	6
*Coloburiscus**	93	0.06	decreaser	0.00000	81.5	6	6
Gyraulus	30	0.14	increaser	0.03756	82	5	5
Copepoda*	25	0.47	increaser	0.16357	80.5	5	5
*Corynoneura**	21	0.55	increaser	0.15439	80.5	5	5
Nematoda*	43	0.69	increaser	0.20152	80.5	5	5
Sphaeriidae*	71	0.37	increaser	0.05772	77.5	5	5
*Paranephrops**	19	0.19	increaser	0.00000	68	5	5
Oxyethira	111	0.06	increaser	0.00482	46.5	3	4
*Paraleptamphopus***	30	0.72	increaser	0.00636	35	2	3
*Tanytarsus**	45	0.28	increaser	0.00358	14.5	1	1

Table 9, continued.

Taxon	No. of occurrences	Model R^2	Response group	Maximum cumulative importance	% Sediment cover threshold at 25% of max. cum. imp.	Tolerance value (raw-scale gradient)	Tolerance value (log-scale gradient)
*Austrosimulium**	165	0.11	increaser	0.00241	13.5	1	1
Potamopyrgus**	228	0.31	increaser	0.00837	13.5	1	1
Orthocladiinae	120	0.09	increaser	0.00463	11.5	1	1
Acarina	26	0.08	unclear	0.00542	80.5		
Aphrophila	111	0.21	unclear	0.00449	17.5		
Austroclima	39	0.42	unclear	0.03683	13.5		
Cladocera	16	0.18	unclear	0.04929	80		
Collembola	29	0.05	unclear	0.00771	44		
Cricotopus	18	0.29	unclear	0.00716	15		
Empididae	24	0.03	unclear	0.00159	16.5		
Eriopterini	112	0.17	unclear	0.01171	11		
Hexatomini	26	0.19	unclear	0.00958	16.5		
Ichthybotus	15	0.45	unclear	0.00000	61.5		
Megaleptoperla	43	0.30	unclear	0.00000	64.5		
Muscidae	50	0.01	unclear	0.00015	17		
Nannochorista	15	0.54	unclear	0.00315	32		
Oeconesidae	21	0.21	unclear	0.00000	63		
Oligochaeta	234	0.43	unclear	0.00570	16		
Ostracoda	84	0.57	unclear	0.06312	33		
Paracalliope	75	0.33	unclear	0.00000	98.5		
Paralimnophila	13	0.21	unclear	0.00117	19		
Physa	67	0.33	unclear	0.01014	77.5		
Platyhelminthes	71	0.07	unclear	0.00000	98		
Polypedilum	25	0.16	unclear	0.00000	68		
Polyplectropus	22	0.15	unclear	0.02427	18.5		
Scirtidae	24	0.50	unclear	0.02154	13		
Tanypodinae	73	0.17	unclear	0.00269	16.5		
Tanytarsini	17	0.27	unclear	0.01264	59		
Triplectides	23	0.20	unclear	0.02089	19		
Zelandobius	65	0.04	unclear	0.00108	14.5		
Zephlebia	34	0.88	unclear	0.01769	8.5		

For the chlorophyll-*a* gradient we identified 20 decreasers and 8 increasers while 37 taxa were classified as 'unclear' (Table 10). Chlorophyll-*a* thresholds ranged from 0.1 to 103 mg chl-*a*/m^2 for decreasers and from 32 to 279 mg chl-*a*/m^2 for increasers. Assignment of tolerance values based on the raw scale of the chlorophyll-*a* gradient resulted in the majority of taxa having a score of 10 (13 out of 20) and few taxa with other scores (Table 10). The assignment based on the log-scale resulted in a somewhat more even distribution across the values of 1–10 although 7 taxa were assigned a score of 6.

Table 10. List of taxa along with model R^2, number of occurrences in the dataset, maximum cumulative importance, and chlorophyll-*a* threshold defined at the point where 25% of the maximum cumulative importance is reached, and tolerance values assigned based on the raw-scale and log-scale chlorophyll-*a* gradient. Note that the models of eight taxa were selected by the R package as having an $R^2 > 0$, however their R^2 values were undistinguishable from zero. * indicates taxa that were decreaser or increaser across the periphyton but not across the sediment gradient, ** indicates taxa that responded to periphyton in the opposite way compared to sediment. Taxa in grey classified as unclear.

Taxon	No. of occurrences	Model R^2	Response group	Maximum cumulative importance	Chl-*a* (mg/m^2) threshold at 25% of max. cum. imp.	Tolerance value (raw-scale gradient)	Tolerance value (log-scale gradient)
Nannochorista*	15	0.54	decreaser	0.10359	0.1	10	10
*Paraleptamphopus***	30	0.72	decreaser	0.04173	0.1	10	10
Scirtidae*	24	0.50	decreaser	0.02750	0.1	10	10
Beraeoptera	22	0.09	decreaser	0.00228	1.9	10	8
Hydraenidae	34	0.28	decreaser	0.01895	1.9	10	8
Zelandoperla	51	0.19	decreaser	0.01306	1.9	10	8
Hexatomini*	26	0.19	decreaser	0.00389	9.4	10	7
Orthopsyche	25	0.50	decreaser	0.03475	9.4	10	7
*Ichthybotus**	15	0.45	decreaser	0.06177	13.2	10	7
*Paralimnophila**	13	0.21	decreaser	0.00475	13.2	10	7
*Zelandobius**	65	0.04	decreaser	0.00213	13.2	10	7
Austroperla	37	0.17	decreaser	0.00703	15.0	10	7
*Zephlebia**	34	0.88	decreaser	0.00439	18.8	10	7
Helicopsyche	28	0.14	decreaser	0.00042	43.1	8	6
Deleatidium	263	0.49	decreaser	0.04726	52.4	8	6
Oeconesidae*	21	0.21	decreaser	0.00000	52.4	8	6
Ameletopsis	10	0.33	decreaser	0.00000	54.3	8	6
Pycnocentrodes	169	0.14	decreaser	0.00002	65.5	7	6
*Potamopyrgus***	228	0.31	decreaser	0.01642	78.6	7	6
Pycnocentria	127	0.12	decreaser	0.01215	102.9	6	6
*Olinga***	119	0.16	increaser	0.00000	278.7	5	5
Gyraulus	30	0.14	increaser	0.00925	220.7	4	5
*Maoridiamesa***	41	0.37	increaser	0.00668	93.6	2	3
*Physa**	67	0.33	increaser	0.02545	80.5	1	3
*Paracalliope**	75	0.33	increaser	0.00917	76.7	1	3
Orthocladiinae	120	0.09	increaser	0.00928	74.9	1	2
Oxyethira	111	0.06	increaser	0.01547	52.4	1	2
Oligochaeta*	234	0.43	increaser	0.02999	31.9	1	1
Acanthophlebia	17	0.74	none	0.04666	16.9		
Acarina	26	0.08	none	0.00098	50.6		
Aoteapsyche	206	0.29	none	0.00117	46.8		
Aphrophila	111	0.21	none	0.01016	43.1		
Archichauliodes	124	0.35	none	0.00116	74.9		
Austroclima	39	0.42	none	0.00649	7.6		
Austrosimulium	165	0.11	none	0.00120	48.7		
Chironomidae	165	0.43	none	0.03101	78.6		
Cladocera	16	0.18	none	0.00000	91.7		
Collembola	29	0.05	none	0.00046	65.5		
Coloburiscus	93	0.06	none	0.00000	325.4		
Copepoda	25	0.47	none	0.00711	41.2		
Corynoneura	21	0.55	none	0.00386	73.0		
Costachorema	33	0.25	none	0.00000	357.2		
Cricotopus	18	0.29	none	0.05964	24.4		

Table 10, continued.

Taxon	No. of occurrences	Model R^2	Response group	Maximum cumulative importance	Chl-*a* (mg/m^2) threshold at 25% of max. cum. imp.	Tolerance value (raw-scale gradient)	Tolerance value (log-scale gradient)
Elmidae	251	0.40	none	0.00662	58.0		
Empididae	24	0.03	none	0.00011	31.9		
Eriopterini	112	0.17	none	0.00584	41.2		
Hydrobiosis	197	0.18	none	0.00545	67.4		
Latia	18	0.03	none	0.00051	9.4		
Megaleptoperla	43	0.30	none	0.00810	7.6		
Muscidae	50	0.01	none	0.00071	104.8		
Nematoda	43	0.69	none	0.01434	93.6		
Nesameletus	56	0.35	none	0.00404	43.1		
Ostracoda	84	0.57	none	0.01661	129.1		
Paranephrops	19	0.19	none	0.00386	13.2		
Platyhelminthes	71	0.07	none	0.00795	52.4		
Polypedilum	25	0.16	none	0.00224	41.2		
Polyplectropus	22	0.15	none	0.00000	321.7		
Psilochorema	140	0.19	none	0.01164	97.3		
Ptilodactylidae	11	0.57	none	0.00000	54.3		
Sphaeriidae	71	0.37	none	0.01757	84.2		
Stenoperla	45	0.18	none	0.00599	97.3		
Tanypodinae	73	0.17	none	0.00939	86.1		
Tanytarsini	17	0.27	none	0.00125	26.3		
Tanytarsus	45	0.28	none	0.00650	13.2		
Triplectides	23	0.20	none	0.00706	13.2		

Out of those 28 taxa responsive to sediment, 11 had unclear response shapes across the periphyton gradient. Out of the 37 taxa responsive to periphyton, 16 had unclear response shapes across the sediment gradient. Four taxa showed opposing response shapes to these two stressors. All these taxa potentially allow stressor-specific metrics to be calculated that are able to discriminate between sediment and nutrient effects.

Metric calculation and preliminary validation

All sediment-specific metrics responded to sediment according to expectations. Sensitive taxon metrics as well as the 'Sediment-MCI' and 'Sediment-QMCI' responded negatively to increasing % sediment cover, while tolerant taxon metrics responded positively (Figure 5). The 'Sediment-MCI' calculated from tolerance values assigned on the raw scale of the sediment cover gradient produced the best linear regression model (R^2 = 0.33) followed by that where tolerance values had been assigned on the log-scale of the sediment cover gradient (R^2 = 0.30). The sensitive taxon metrics performed less well (R^2 = 0.22) and sediment cover only explained a small proportion in the variation of the tolerant taxon metrics (R^2 = 0.12) (Figure 5).

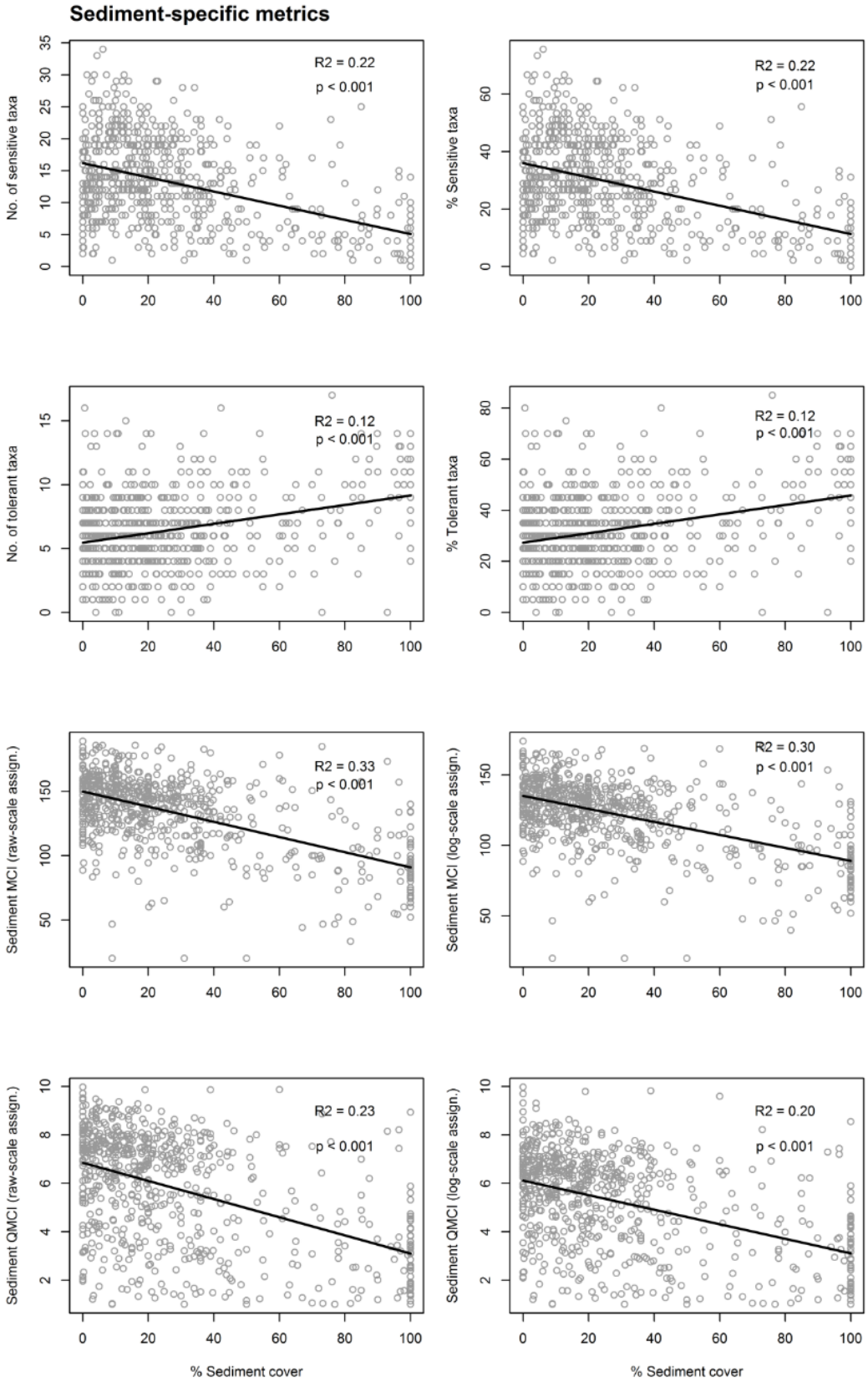


Figure 5. Sediment-specific metric responses to % fine sediment cover as modelled by linear regression analysis (p-values and R^2 are also provided).

Among the nutrient-specific metrics, the sensitive taxon metrics did not respond to chlorophyll-*a*, but all others responded according to expectations (Figure 6). 'Nutrient-MCI' and 'Nutrient-QMCI' responded negatively to increasing chlorophyll-*a*, while tolerant taxon metrics responded positively. However, the nutrient-specific metrics performed overall far less well than the sediment-specific metrics. The best metrics were the 'Sediment-MCI' and 'Sediment-QMCI' versions, all of which had an R^2 of about 0.10 (Figure 6).

We also plotted the sediment-specific metrics across the chlorophyll-*a* gradient and the nutrient-specific metrics across the sediment cover gradient to test whether the metrics have the potential to discriminate between the effects of these two stressors. Sediment-specific metrics predominantly did not respond to increasing chlorophyll-*a* and the ones that did, the tolerant taxon metrics, had very low R^2 values (R^2 = 0.03, Figure 7). By contrast, all nutrient-specific metrics also responded to increasing sediment cover although the relationships were not as strong (R^2 up to 0.22, Figure 8) as those for the sediment-specific metrics with sediment cover (Figure 5).

Finally, while there was a limited amount of indicator taxa (decreasers and increasers) that we were able to identify using this approach, they were present in all of the 775 samples, except for 4 samples for sediment and 8 samples for nutrient (decreasers or increasers) indicator taxa.

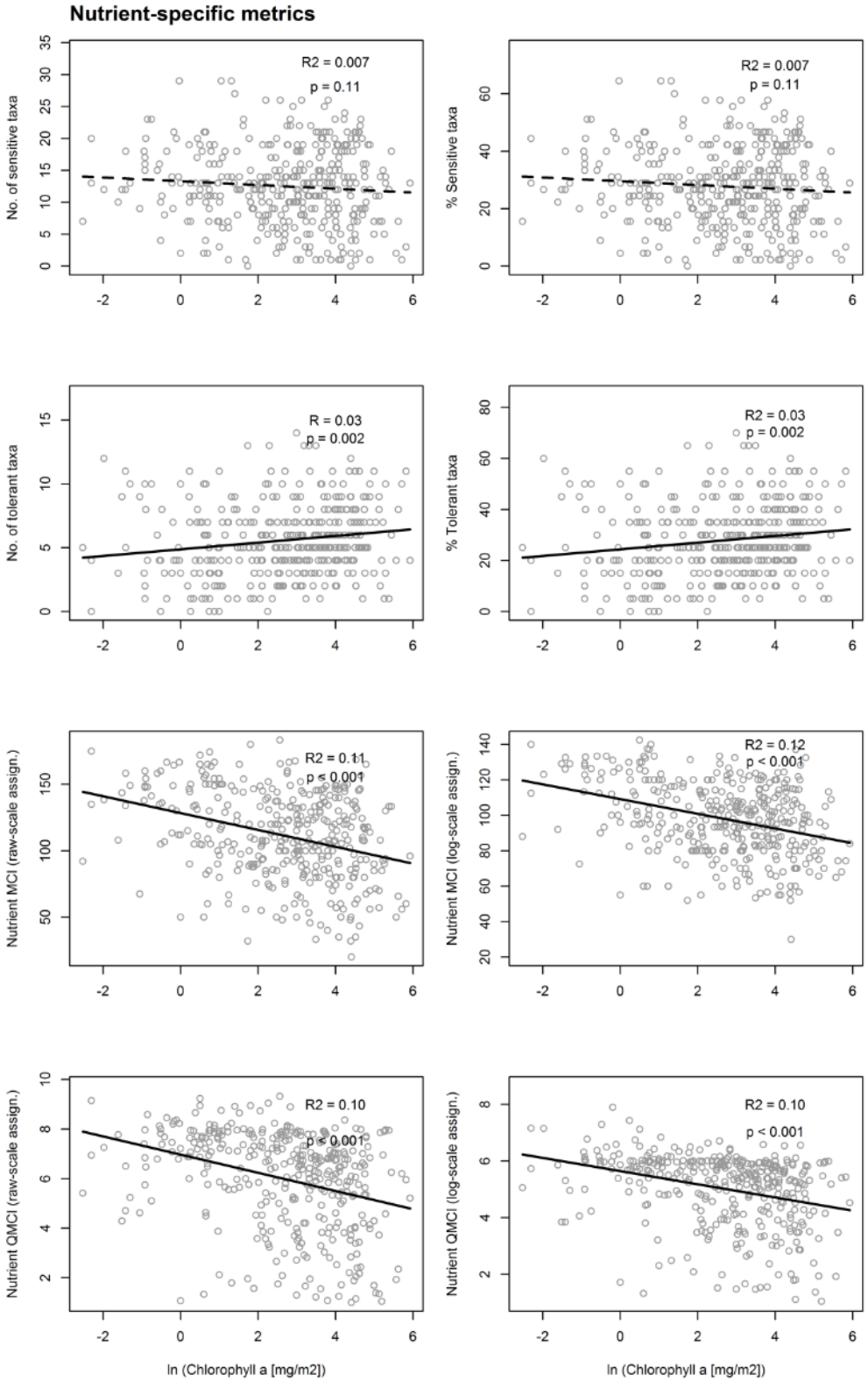


Figure 6. Nutrient-specific metric responses to chlorophyll-*a*. (p-values and R^2 are also provided). Dashed lines indicate that the linear regression model was not statistically significant at p = 0.05.

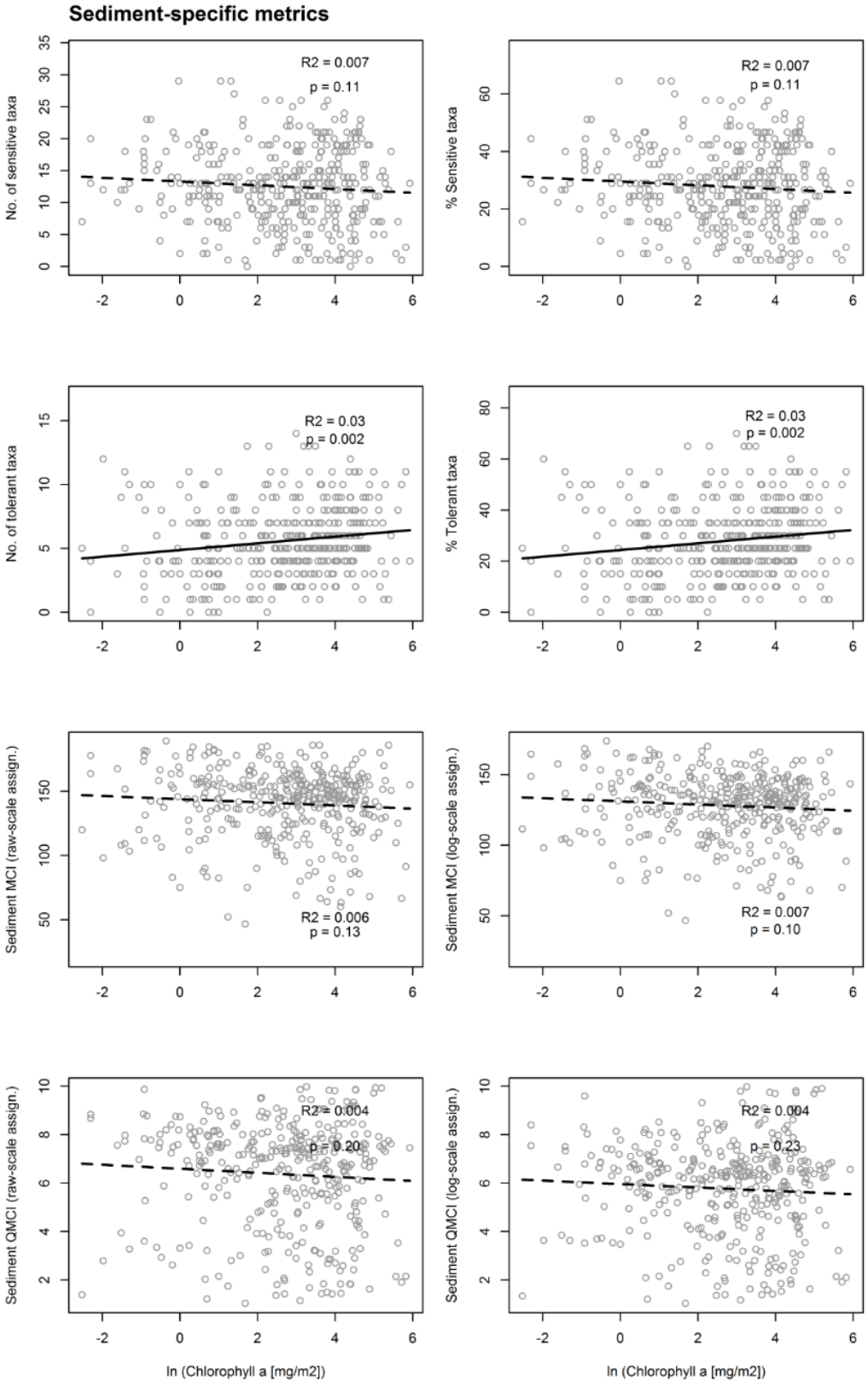


Figure 7. Sediment-specific metric responses to chlorophyll-*a* as modelled by linear regression analysis (p-values and R^2 are also provided). Dashed lines indicate that the linear regression model was not statistically significant at p = 0.05.

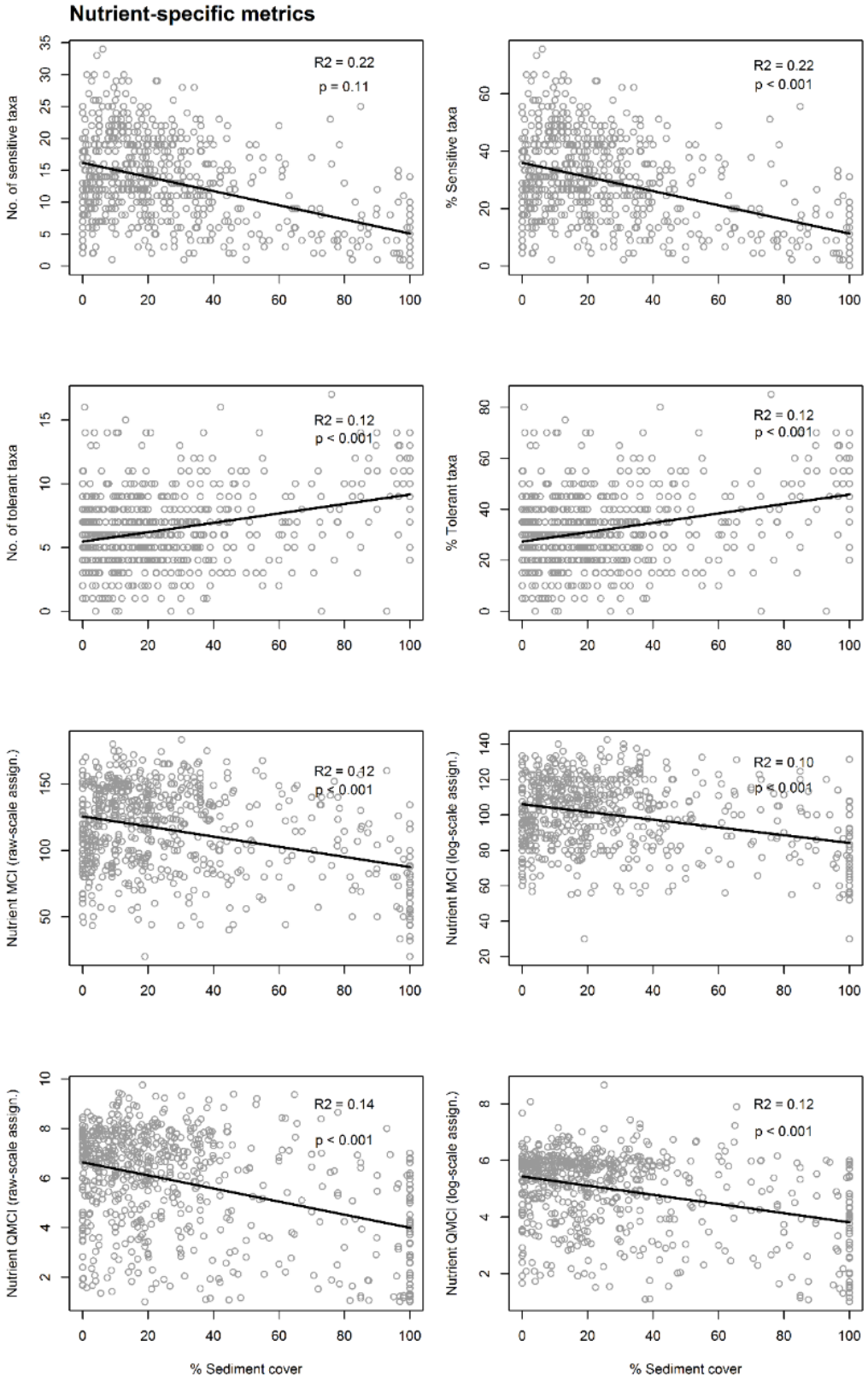


Figure 8. Nutrient-specific metric responses to % fine sediment cover as modelled by linear regression analysis (p-values and R^2 are also provided). Dashed lines indicate that the linear regression model was not statistically significant at p = 0.05.

3.5.4. Discussion and recommendations

Overall, our analysis suggests that stressor-specific metrics can be developed because we identified a relatively high proportion of taxa (among the decreasers and increasers) whose relative abundance was responsive to sediment but not responsive to periphyton (indicated by chlorophyll-*a*). Preliminary validation using the training data set indicates that sediment-specific metrics may have the potential to be able to discriminate between sediment and nutrient effects, and hence be used as diagnostic tools. Nutrient-specific metrics, on the other hand, may be less useful.

However, only validation of the metrics using an independent dataset will show whether these metrics have diagnostic power and relationships to their respective stressor gradients strong enough to be useful in a policy, management or regulatory context. Such datasets are already available. First, we only used a subset of the research data to develop metrics and other research data are still available for validation. In particular, data from manipulative experiments designed to test causal relationships should be very suitable to test diagnostic power. Furthermore, the large nationwide SoE dataset could also be used for validation.

In this project, our stressor-specific metrics are linked to stressors and compared with other existing metrics in another task (see Section 5). This will show whether there is any gain in the further development of such metrics.

In this sub-task, we conducted statistical analyses to develop tolerance values to calculate new stressor-specific indices. While validation is a must for any new metric, there is also scope for future work which could investigate alternative options using the gradient forest approach to refine threshold definition and tolerance value assignment as well as alternative options for metric calculation to improve the performance of stressor-specific metrics. This work would be done alongside validation as the validation exercise will help with deciding which metrics are among the best. We recommend further work to investigate:

1. taking into account the strength of the relationships between the taxa and the stressor to inform tolerance values
2. the inclusion of taxa classified as unclear in metric calculation including those that showed subsidy-stress responses
3. the exclusion of taxa in metric calculations that had very small maximum cumulative importance and/or taxa that have very small R^2 values (requires setting of some criterion)
4. the use of presence-absence macroinvertebrate data or density data instead of relative abundance data
5. the use of expert opinion to assign tolerance values to taxa that were not present in the dataset or did not make it into the analysis as they had less than 10 occurrences.

We also recommend investigating other approaches to tolerance value assignment such as the approach used for the MCI_{hb} but using individual, measured stressor gradients rather than an *a priori* overall pollution gradient based on expert opinion or the approach used for the MCI_{sb} (i.e. an iterative rank correlation procedure following the method by Chessman (2003).

4. EXPLORATION OF A MULTIVARIATE APPROACH

4.1. Overview

Multivariate predictive models are used globally to provide a reference condition based assessment of stream communities, but a national macroinvertebrate model is yet to be developed for New Zealand. This aim of this task was to develop a nationwide multivariate predictive model for stream macroinvertebrates in New Zealand. This section describes a first step undertaken working towards this aim, specifically a test of two differing approaches to analysing taxa at reference sites to predict the probability of taxa occurrence at test sites:

1. an existing stream typology (FENZ), and
2. a new biological classification based on the relationship between taxa and environmental descriptors.

A reference site dataset was compiled (described in Section 2.2.3) and analysed using a RIVPACS approach. Probability of occurrence of taxa was used to calculate common macroinvertebrates metrics and provide proof-of-concept of a national multivariate model.

4.2. Introduction

A major challenge for bioassessment is ensuring that any index provides consistent meaning in different environmental settings; that is, a given score from an index should indicate the same biological condition irrespective of geographic location or stream type (Mazor et al. 2016). Therefore, effective bioassessment indices should account for naturally occurring variation in aquatic assemblages so that deviations from reference conditions resulting from anthropogenic disturbance are minimally confounded by natural variability (Hughes et al. 1986; Reynoldson et al. 1997). This is particularly important where bioassessment is used in regulatory frameworks, such as the NPS-FM, as errors in assessment can lead to management actions with significant financial and resource implications.

This issue is commonly addressed in bioassessment through the use of the reference condition approach (Reynoldson et al. 1997). This approach distinguishes natural variability from anthropogenic impacts by comparing the biological attributes from a test site with a group of similar sites that are in a minimally-disturbed reference condition.

The need for a baseline for stream macroinvertebrate-based bioassessment was described by Armitage et al. (1983), such that the natural variability in macroinvertebrate communities required site-specific target values against which to

compare results from individual sites. The concept has become central to many bioassessment programmes, where the target is referred to as the 'expected' (E) value and the result from a test site is referred to as the 'observed' (O) value. The ratio (O/E) of these values is now commonly used as a basis for bioassessment, with values close to 1 indicating sites close to reference condition and values closer to 0 indicating impaired sites. Importantly, the concept of O/E can be applied to any measure or metric of the stream macroinvertebrate community, or any other indicator of interest.

There have been two broad approaches for predicting the reference condition (E) for use in bioassessment programmes, with the key difference being how the reference sites are used to support the prediction of the expected fauna (E). These approaches have been termed 'multivariate' and 'multimetric' by Reynoldson et al. (1997), however this terminology is problematic as these terms are not mutually exclusive in the reference condition approach to bioassessment (for example, multi-metric indices can be developed using multivariate models). A better terminology describes the way in which reference sites can be used to predict the expected fauna, either predictions based on an 'environmental' classification of stream types or a model based on multivariate 'biological' data that uses environmental attributes to predict the macroinvertebrate community.

The biological (previously 'multivariate') approach is based on a biological classification of reference sites. The multivariate community composition of reference sites is determined and a classification algorithm applied to assign classes. Then, the relationship between this classification and the environmental characteristics of the sites is used to predict the appropriate reference condition against which to compare a test site (Clarke et al. 2003). Hence, in the biological approach, the taxonomic composition of sites is predicted using a model driven by environmental attributes. Here the rationale is that the biological communities reflect the physiochemical environment, but there is no prior assumption as to which environmental features account for the biological variation.

The environmental (previously 'multi-metric') approach is based on an environmental classification of reference sites. The environmental character of reference sites is determined (e.g. geomorphology, physicochemistry, flow) and a classification algorithm applied to assign classes. This approach is based on the concept that the biological properties of aquatic ecosystems can be inferred from environmental knowledge of the region in which they occur (Hawkins & Norris 2000). The environmental approach does not predict taxonomic composition of sites, rather metrics of the community based on an *a priori* classification. This approach involves assumptions about how macroinvertebrate communities are structured along environmental gradients.

The predictions from the two approaches also differ in that the environmental classification approach involves type-specific predictions—all sites within a particular environmental class are expected to have the same macroinvertebrate fauna and metric scores under reference condition. The biological classification approach offers unique site-specific predictions that are based on the probability of a test site belonging to all possible biological classification groups.

The biological approach has been trialled previously at regional scales in New Zealand (e.g. Waikato (Coysh & Norris 1999; Death & Collier 2010) and Manawatu-Wanganui (Joy & Death 2003)), however it is yet to be applied at a national scale. The aim of this task was to attempt to develop a nationwide biological model based on multivariate community composition for the prediction of site-specific macroinvertebrate fauna at reference condition. We describe the relative performance of environmental and biological classification approaches for predicting macroinvertebrate reference condition and compares these approaches with a null model (i.e. a model that makes no attempt to explain variability in communities; van Sickle et al. 2005).

4.3. Methods

4.3.1. Dataset

We used a reference dataset from 538 sites draining catchments with greater than 80% native vegetation upstream (Section 2.2.3). Our selection of 80% native vegetation cover was a rational decision to include as many ‘minimally-disturbed’ sites as possible recognising that it is important that any dataset used to construct a predictive model should represent the full range of conditions where it is to be used (Mazor et al. 2016). We did not take other land use criteria into account as other studies have (e.g. Ode et al. 2016). The 538 sites have broad geographic coverage of the country (Figure 9), but there are gaps which coincide with intensively used land (i.e. where reference conditions are unlikely to be found). All sites were used in model development and a single sample was randomly chosen from each site when multiple samples were available.

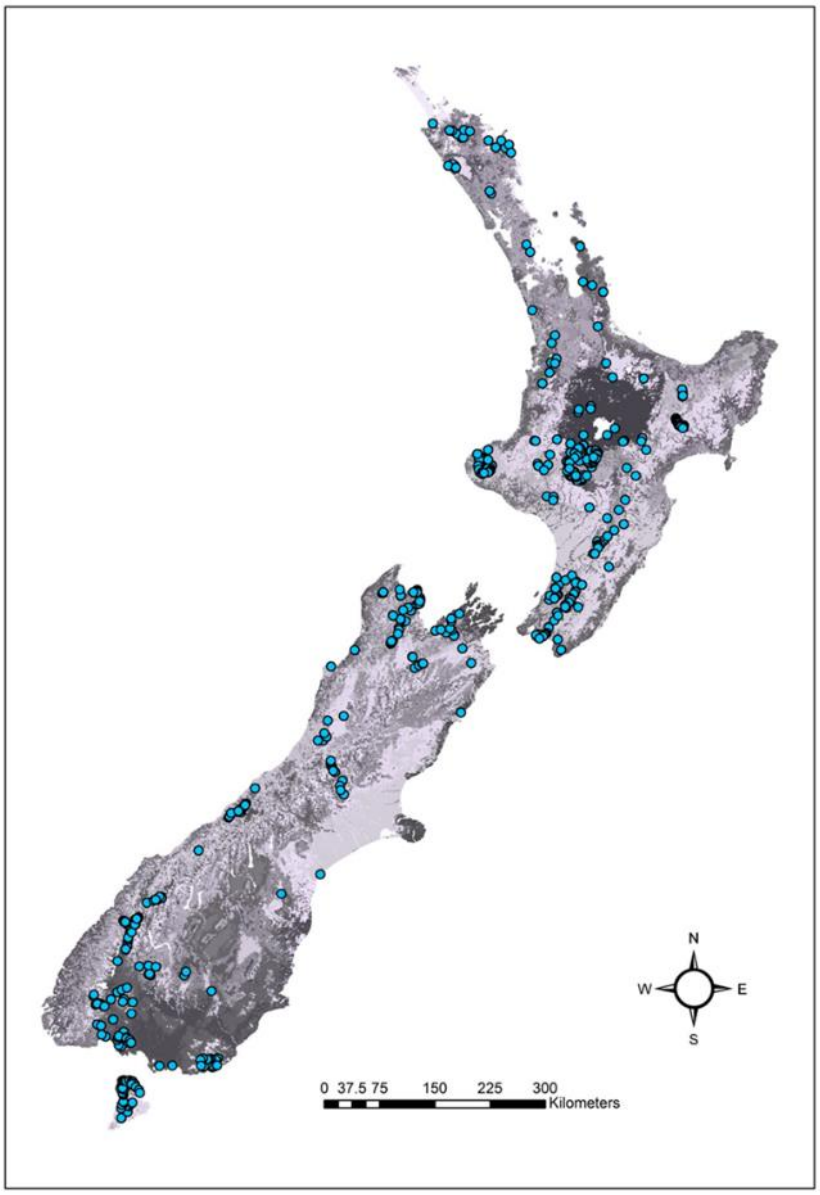


Figure 9. Location of 538 sites with > 80% native vegetation and source of macroinvertebrate data for the multivariate analyses.

4.3.2. Classifications

Predictive models were constructed using the different site classification (biological and environmental) approaches and used to predict the expected fauna as follows.

Environmental classification

In New Zealand there are two pre-existing river classifications, the River Environment Classification (REC (Snelder et al. 2004)) and the Freshwater Ecosystems of New Zealand (FENZ (Leathwick et al. 2010; 2011).) classification. They differ in what variables were used to develop their classifications, but both are based on catchment-

scale and segment-scale descriptors of the stream environment (e.g. geology, stream flow/source of flow, climate). The FENZ classification however was further biologically-optimised during development using benthic macroinvertebrate and fish community data to ensure river types represented distinct biological communities (this was not done in the REC) (Leathwick et al. 2010; 2011). Rather than redefine an environmental classification based on the physical qualities of the 538-site dataset, we chose to use the FENZ classification.

We used the FENZ Level 2 class (100 group). The expected macroinvertebrate metrics (e.g. number of taxa and MCI) for a test site were determined as the mean of the observed metric values for all reference sites in the FENZ group to which the test site is assigned.

Biological classification

The predictions based on the biological classification of sites were made by constructing a River Invertebrate Prediction and Classification System (RIVPACS)-type model using the framework described by Clarke et al. (2003) although we utilised more recently developed statistical tools where appropriate. The model was constructed in R using default settings in scripts provided by John van Sickle (US Environmental Protection Agency).

First, we developed a biological classification of the reference site data using agglomerative nesting (agnes) routine (Kaufman & Rousseeuw 1990) in the R package 'cluster' based on the Sorenson dissimilarity index. We created a dendrogram to visualise this classification and select the end groups to use in the model development.

Second, we constructed a 500-tree random-forest model (Cutler et al. 2007) using the randomForest package in R (Liaw & Wiener 2002) to predict group membership for sites. Potential environmental predictor variables were assembled from a number of sources, including FENZ, REC and unpublished datasets[5]. Predictor variables were only considered for inclusion in the model if they were considered to have a causal relationship with the distribution of macroinvertebrates and were not likely to be substantially affected by human activity. All predictors were used in an initial random-forest model to identify those predictors that were important for predicting sites into biological groups as measured by decrease in accuracy and the Gini index (Liaw & Wiener 2002). The final model was based on the subset of predictors that were most important in the initial model run.

Third, the probability of each taxa occurring at a site was determined based on the probability of group membership for the test site and the frequency of occurrence of each taxon within each group. The probability of occurrence for a taxon at a test site

[5] MfE provided spatial predictors of sediment (author: Joanne Clapcott, Cawthron) and stream flow (author: Doug Booker, NIWA).

was calculated as the weighted average of the frequencies of occurrence of a taxon across groups in which frequencies are weighted by the probabilities of group membership.

The number of taxa expected at a site was determined as the sum of all probabilities for that site. The expected Macroinvertebrate Community Index (MCI; Stark 1985) at each site was calculated from the probability of each taxon's occurrence at a site using the method described by Clarke et al. (1996). This approach was developed for calculating Biological Monitoring Working Party (BMWP; Hawkes 1998) scores expressed as Average Score Per Taxon (ASPT) using the RIVPACS approach in the UK. Given that the MCI scoring system is based on the BMWP system and that the MCI score is calculated in the same manner as the ASPT, this is considered an appropriate method to calculate expected MCI scores.

Null model

In the null model the predicted metrics for sites are the average observed value for all reference sites (van Sickle et al. 2005). The null model makes no attempt to explain the variation in taxa occurrences among the reference sites and as a result provides an upper limit for standard deviation (SD) (O/E), a limit that would be achieved if a predictive model failed to account for any of the variation in macroinvertebrate communities.

4.3.3. Model assessment

We used the approach described by van Sickle et al. (2005) to investigate the predictive accuracy of models based on the differing classification approaches. In a useful model, the observed (O) fauna for reference sites should closely resemble the expected (E) fauna based on the model's predictions. Therefore, the O/E ratio for a metric should vary around unity and the standard deviation (SD) of the O/E ratio for a metric should be low. In this study, the assessment of predictive accuracy was based on the O/E values of two metrics, number of taxa and MCI.

The SD (O/E) of each of the environmental and biological classification approaches were compared directly to assess the accuracy of the faunal predictions. Additionally, the SD (O/E) of each model was compared with that of a null model. In addition, for the environmental and biological models we calculated the percent reduction in SD (O/E) compared with the null model as the percent reduction provides a standardised measure of performance that can be used to compare the performance of models developed in different locations.

4.4. Results

4.4.1. Environmental classification

Fifty FENZ classes, including all the most abundant classes, were represented in the reference data (Figure 10). There were between 1 and 84 sites per FENZ class with an average of 11 sites per class. The mean O/E for the sites was 1.00 with a SD of 0.22 (Table 13).

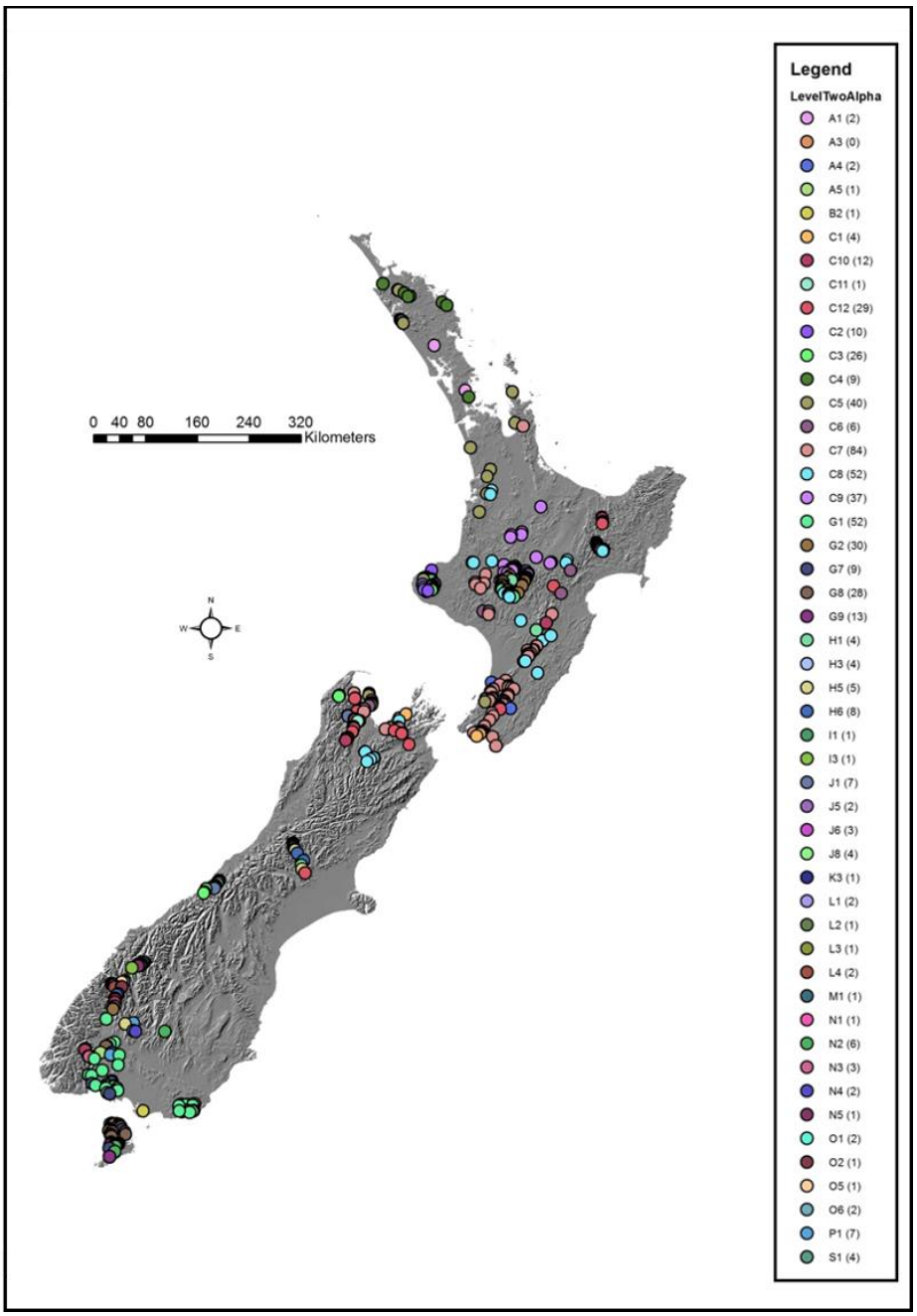


Figure 10. Sites coloured by various Freshwater Ecosystems of New Zealand classifications.

4.4.2. Biological classification

For the purposes of the biological model, the 538 sites were classified into 13 groups based on the macroinvertebrate community data (Figure 11). These groups were determined based on a visual assessment of the dendrogram. The groups varied in size from 13 to 83 sites (Table 11). There is evidence of inter-group variability in the environmental characteristics for each group (Figure 12), which indicated the potential to predict group membership based on these parameters.

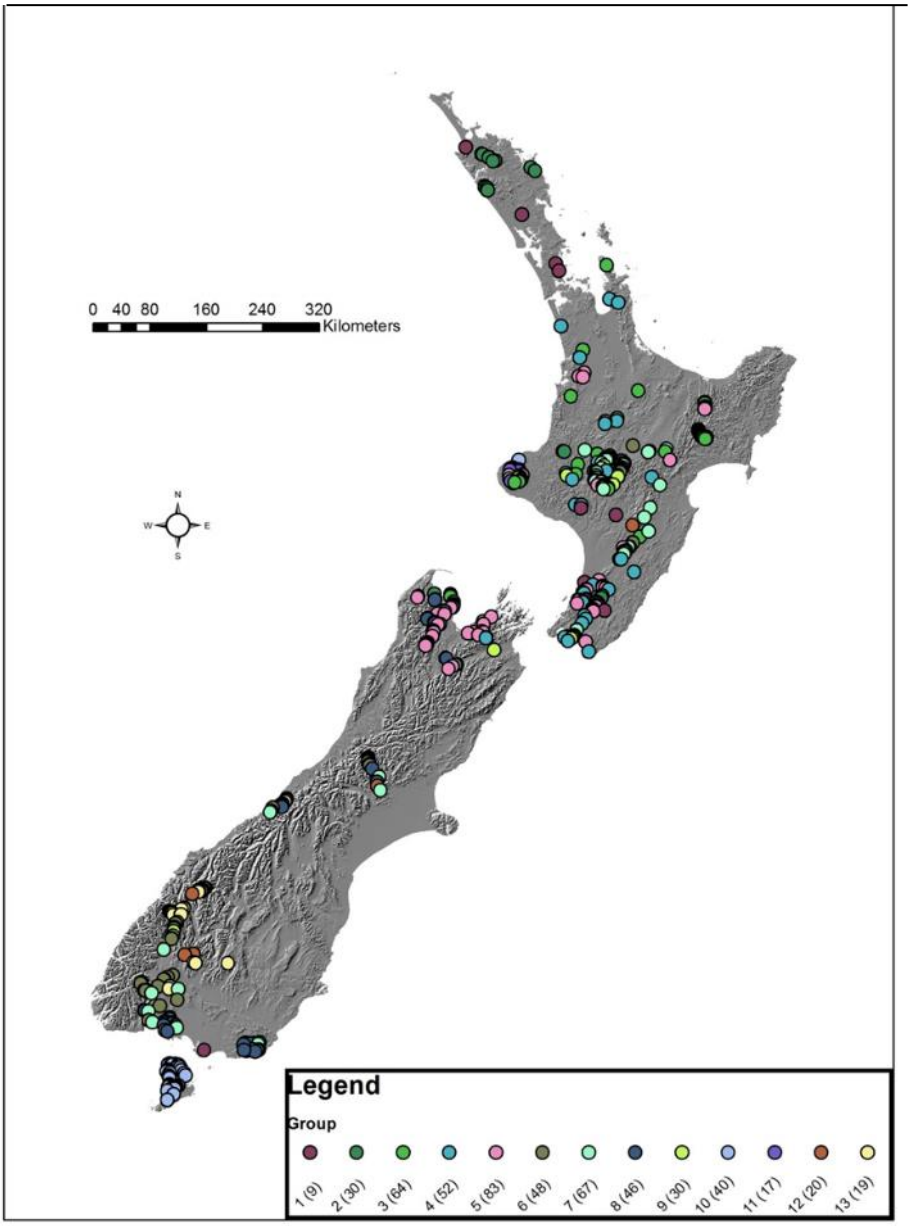


Figure 11. Sites showing different biological classification groups based on macroinvertebrate community data.

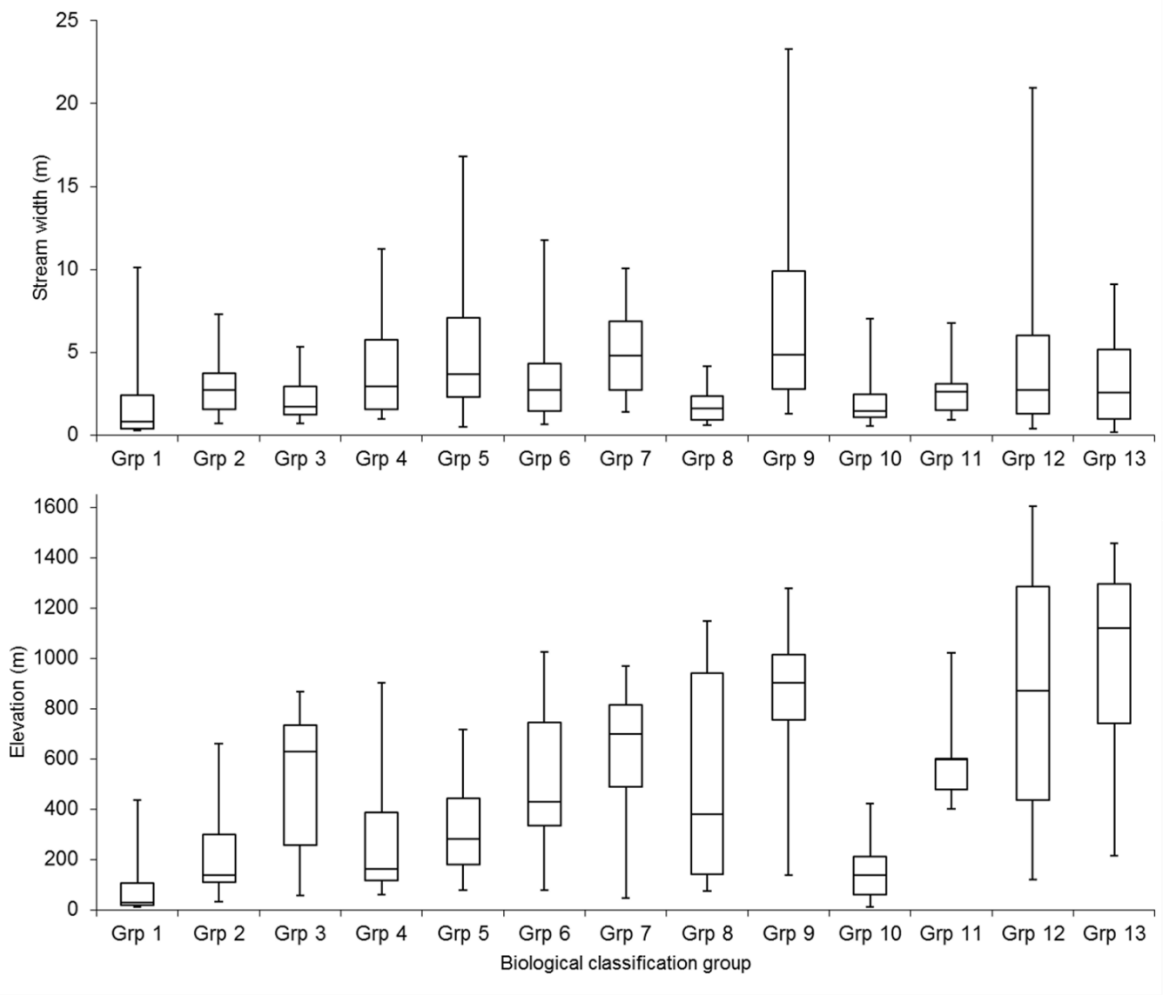


Figure 12. Intra- and inter-group variability observed in stream width and elevation for groupings defined by the biological classification.

Table 11. Biological group sizes and mean environmental conditions. See Table 12 for explanation of environmental variables.

Group	Number of sites	Easting	Northing	SegJanAirT	SegMinTNorm	DSDist2Coast	USDaysRain	USAvgSlope	USHardness	Elevation
1	13	2381306	6190563	17.8	1.4	42	8.8	9.9	2.68	107
2	33	2546360	6333014	17.5	1	81	17.9	17.2	3.57	234
3	69	2670587	6199026	15.9	0.8	104	23.8	15.5	2.89	523
4	54	2648402	6129893	16.4	0.9	90	19.6	20.7	3.66	299
5	83	2598553	6051013	15.9	0.4	77	27.1	22.9	3.57	338
6	48	2280223	5698657	13.7	0.9	107	28.9	22.5	3.68	516
7	67	2620862	6064966	14.6	0.6	176	23.5	13.2	2.79	628
8	46	2366270	5733610	13.5	1.3	87	18.9	16.2	3.67	533
9	30	2535449	5997608	13.6	0.6	231	43.3	20.5	3.65	832
10	40	2127445	5379935	12.4	3.3	6	10.4	15.5	4.41	167
11	17	2597897	6217247	15	2.5	22	48.4	12.1	2.19	599
12	20	2300803	5726023	12.3	1.3	169	45	27.8	4.08	862
13	19	2146595	5571420	11.1	1.4	156	42.1	29	4.02	971

4.4.3. Biological model predicting probability of taxa occurrence

An initial random forest model was developed using the 18 potential environmental predictors identified during our selection process (Table 12). Based on the mean decrease in accuracy of these predictors in the initial model run (Figure 13), nine predictors were retained in the final model (Table 12 and Figure 14).

Table 12. Environmental variables used as predictors in a random forest model (including data source). All variables were used in the original model. The subset indicated in column 3 was retained in the final model.

Variable	Description	Used in final model
Easting		✓
Northing		✓
Elevation	Metres a.s.l.	✓
SegJanAirT	Summer (January) air temperature (degrees C)—used in the absence of robust estimates of water temperature (FENZ)	✓
SegMinTNorm	Average minimum daily air temperature (degrees C) normalised with respect to SegJanAirT—negative values indicate strongly seasonal climates and positive values indicate weakly seasonal climates (FENZ)	✓
USDaysRain	Days/year with rainfall greater than 25 mm in the upstream catchment (FENZ)	✓
USAvgSlope	Average slope in the upstream catchment (degrees), describes catchment-driven modification of flow variability (FENZ)	✓
DSDist2Coast	Distance to coast (km), from mid-point of each river segment (FENZ)	✓
USHardness	Average hardness (induration) of surface rocks using values derived from the underlying LENZ layers (FENZ)	✓
Order	Strahler stream order	
USAvgTNorm	Average air temperature (degrees C) in the upstream catchment, normalised with respect to SegJanAirT, with negative values indicating colder (higher elevation) headwaters than average, given the segment temperature, and positive values indicating warmer temperatures (FENZ)	
USArea	Catchment area upstream of each location (REC)	
SegFlow	Mean annual flow (m^3/sec), derived from hydrological models (FENZ)	
WidthMALF	Wetted width across the river channel (m) at mean annual low flow. Lower values are less wide (Booker 2010)	
SEDE	Predicted expected percentage fine sediment cover under reference state (Clapcott et al. 2011)	
SegSlope	Segment slope (degrees), derived from GIS calculation using length and difference between upstream and downstream elevation for each segment (FENZ)	
ReachSed	Weighted average of proportional cover of bed sediment using categories of: 1–mud; 2–sand; 3–fine gravel; 4–coarse gravel; 5–cobble; 6–boulder; 7-bedrock, predicted from a boosted regression tree model (FENZ)	
SegSlopeSqrt	Square-root transformed segment slope (slope +1) (FENZ)	

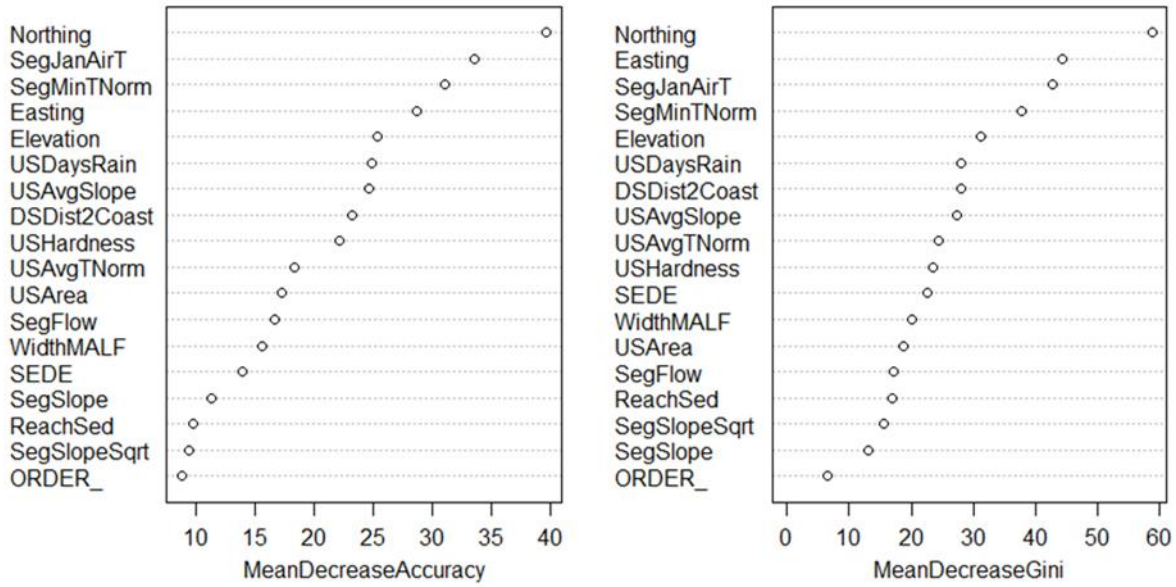


Figure 13. Relative importance of environmental predictors in an initial random forest model (high Gini index = higher importance for that predictor and a greater reduction in model accuracy).

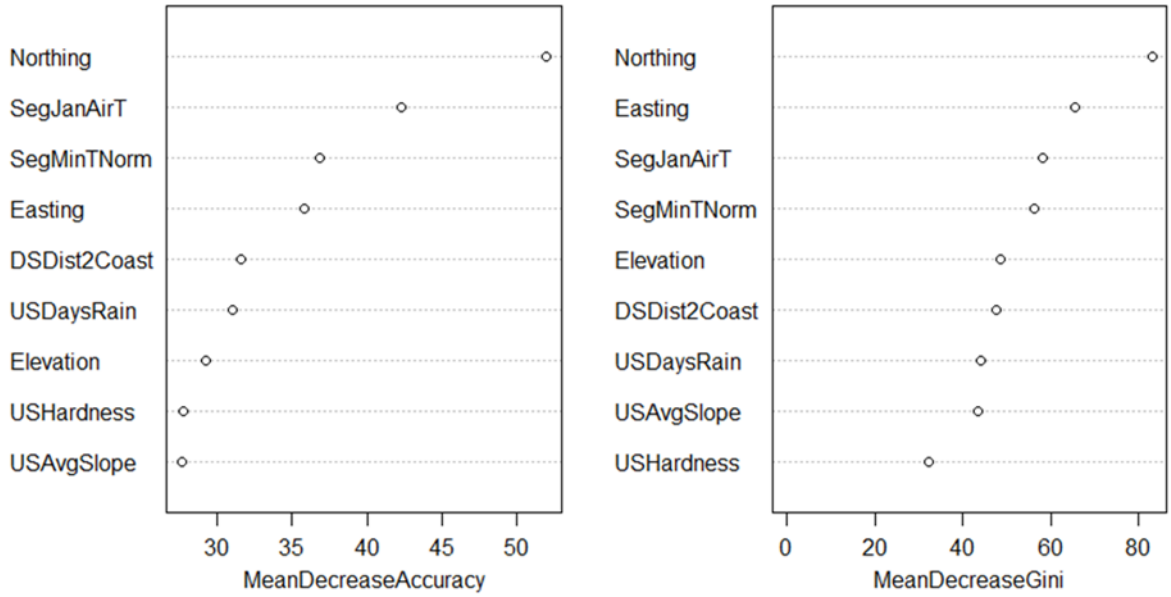


Figure 14. Relative importance of environmental predictors in a final random forest model.

The error rate for group classification was marginally worse for the final model (48%) compared with the initial model (47%). The error rate refers to the proportion of sites that are predicted to the correct group by the model (indicated by the highest probability of group membership); whilst it is important to minimise this error rate, it is

not critical for model performance that a site is predicted to the correct group. This is because the model produces a probability of each site belonging to all groups and the macroinvertebrate community predictions are based on all of these probabilities. Where model performance is not significantly reduced but it is recommended to use smaller sets of environmental predictors (John van Sickle, Oregon State University, pers. comm.). As such, we used the reduced model to predict the probability of taxa occurrence.

4.4.4. Predictive accuracy

The use of an environmental or biological classification resulted in a reduction of the SD (O/E) for both test metrics, and hence an increase in predictive accuracy, when compared with the null model (Table 13).

Table 13. Predictive accuracy of models shown by the standard deviation (SD) of O/E for test metrics and % reduction compared with null model in parentheses.

Model	SD (O/E) Number of taxa	SD (O/E) MCI
Multivariate (biological)	0.214 (14%)	0.088 (19%)
Multimetric (environmental)	0.220 (12%)	0.084 (22%)
Null	0.249	0.108

In all of the models the SD (O/E) was lower for MCI than for Number of taxa, as has been found in similar evaluative studies elsewhere (Moss et al. 1999; van Sickle et al. 2005; Davy-Bowker et al. 2006; Neale & Rippey 2008). For Number of taxa, the biological model produced the lowest SD (O/E) resulting in a reduction of 14% for Number of taxa when compared with the null model. For MCI, the environmental model produced the lowest SD (O/E) resulting in a reduction of 22% for MCI when compared with the null model (Table 13).

4.5. Discussion

This task represents the first attempt to develop a nationwide multivariate biological model for stream macroinvertebrates in New Zealand. As with similar studies elsewhere, the predictive accuracy of a multivariate biological model was greater than that of a null model. The percentage reductions in SD (O/E) achieved using the multivariate model in New Zealand compared with the null model (14% (number of taxa) and 19% (MCI)) are within the range of reductions observed in similar studies elsewhere (12% to 34% in European rivers (Davy-Bowker et al. 2006); 14% to 53% in

US rivers (van Sickle et al. 2005); 22% to 31% for lakes in Ireland (Neale & Rippey 2008).

The comparability in performance of the model reported here with the international examples referenced above is encouraging, particularly given some of the European studies were multi-year programmes set up specifically to develop predictive models. In contrast, the model developed here is considered a 'proof of concept' demonstration of the multivariate biological approach—this model was based on a dataset compiled opportunistically from numerous sources, rather than a dataset explicitly designed and collected to support a predictive model. In addition, there remains scope to refine the different components of the model to increase its performance further.

An assessment of the performance of the biological and environmental models indicates similar results; the biological model results in slightly better predictions of number of taxa, whereas the environmental model results in better predictions of MCI. The relative performance of the environmental model is surprising as previous studies have indicated that environmental models have had limited ability to support predictive models (Reynoldson et al. 1997; Gerritsen et al. 2000; Hawkins & Vinson 2000; Heino et al. 2002; Parsons et al. 2003). Hawkins et al. (2000) hypothesised that one of the reasons for the poor performance of environmental classifications in predicting macroinvertebrate faunas is that biologically important environmental heterogeneity may exist among sites that is not accounted for by broad category environmental classifications. Some of the FENZ categories are large, which potentially results in considerable biological variation within each type. As a result they concluded that, used alone, environmental classifications lead to imprecise predictions of the expected biota at test sites. The results of this study provide limited support for this conclusion. The biological optimisation process used in FENZ may be an important differentiator from previous environmental classification approaches. Conversely, there is much scope for improvement in the performance of the biological approach (see recommendations), whereas the environmental approach may be close to optimal performance.

It should also be recognised that the environmental data on which FENZ is based have been invaluable in the development of the multivariate biological model. These data have been used as predictor variables in the model predicting probability of taxa occurrence. They have been used previously to predict probability of fauna occurrence for all stream segments in the digital river network of New Zealand (Leathwick et al. 2008, 2009). Building the biological model in this way offers the possibility of being able to carry out desktop predictions of the macroinvertebrate fauna for all rivers in New Zealand where they are represented in FENZ, whereas the initial RIVPACS-type models required the collection of field-based environmental data to make predictions of the macroinvertebrate fauna at a test site.

The implications of an imprecise and inaccurate prediction model are twofold. First, the accurate assessment and reporting of the ecological status of a test site is dependent on the faunal predictions from the model. Second, and more importantly, misclassification of a test site by the model could result in a programme of management and conservation activities directed towards an inappropriate ecological target.

4.5.1. Conclusion

The scope of this task was an exploration of the multivariate approach to provide proof of concept, and results have indicated that a multivariate biological model is feasible. The approach was able to produce useful macroinvertebrate community predictions and is therefore likely to be a valuable addition for managing the ecological health of rivers in New Zealand. The ability to use desktop-based environmental predictors is a big advantage over earlier predictive models and offers the potential to predict the macroinvertebrate community for every river reach in New Zealand without having to visit each location (e.g. Hill et al. in press).

This work represents the early development of a predictive model that could provide a robust basis for assessing macroinvertebrate communities in New Zealand rivers. The model accounts for the natural variability in macroinvertebrate communities and therefore offers the potential to monitor and manage our rivers with a nationally consistent suite of indices based on deviation from reference condition. Using the MCI as an example, the model provides site-specific predictions based on environmental conditions at that site, and therefore makes the requirement for habitat or region specific variants of the MCI redundant. Such an approach would support the use of a nationally consistent approach to using macroinvertebrates in the NPS-FM framework.

Whilst this model performed well at a similar level of accuracy to international examples, we recommend that further work is required before such a model is used in an operational manner. This work would be aimed at refining the model to improve the accuracy of predictions and broadening the utility of the model beyond the test metrics used in this demonstration. There remains much scope for this refinement; specific tasks include the need to:

- source additional reference site data to increase coverage of the environmental variability in New Zealand and provide independent validation data
- evaluate the effect of species abundance thresholds for inclusion in model development and predictions
- explore alternative environmental (e.g. REC vs FENZ) and biological classifications (different multivariate classifications) and test their strength
- iteratively test different combinations of environmental predictors to find the optimal set for predicting taxa occurrence

- use the macroinvertebrate predictions to develop site-specific targets for other macroinvertebrate community metrics that are used in New Zealand (e.g. EPT) or proposed elsewhere in this report (e.g. stressor specific metrics, species trait metrics)
- explore the use of abundance data to allow the prediction of abundance-based metrics
- test the performance of the model with test sites of known impact to understand response gradients.

5. LINKING MACROINVERTEBRATE METRICS TO STRESSORS

5.1. Overview

The main aim of this task was to quantify the relationship between metrics and land use and proximate stressors to:

1. Select the most suitable metrics for inclusion in an ecosystem health assessment framework (Section 6). For a macroinvertebrate metric to be used as an indicator for ecosystem health it may not necessarily respond to specific stressors and instead may indicate the cumulative effects of multiple stressors resulting from human land use.
2. Identify whether any macroinvertebrate metrics are predominately linked to a specific stressor to help inform management actions such as the setting of limits on stressor loads, e.g. sediment or nutrients.

New and existing metrics, described in Section 3.5 and Section 2 respectively, were analysed using two statistical techniques which partition the deviance explained by stressors in the presence of environmental factors. Results quantify how much variation can be attributed to the stressors and environmental covariates.

5.2. Introduction

Metrics indicative of general stream health respond to gradients in human land use intensity, such as urbanisation and agricultural land uses, which have multiple pathways through which they affect ecosystem health (refer Figure 1 on page 2). By contrast, stressor-specific metrics should mainly respond to a single stressor gradient. Major land use-derived stressors in streams in New Zealand (and worldwide) include the addition of excess nutrients and fine sediment as well as flow and temperature alteration. Deposited sediment in particular is considered a ‘master stressor’ because effects are often negative in their own right, and interactions with other stressors can make these effects even worse (Townsend et al. 2008; Wagenhoff et al. 2011).

In this section we explore the quantitative link between macroinvertebrate metrics and catchment and reach-scale drivers of nutrient enrichment and deposited sediment. We applied two different statistical techniques, (1) gradient forest analysis, and (2) multiple linear regression models. Both model approaches take into account multiple predictors and hence help with teasing apart the relative importance of multiple stressors and other environmental variables. The gradient forest approach was used as an exploratory tool and provided insight into the shape and strength of the stressor-response relationship of each metric in relation to the other metrics. It was used to reduce the large number of macroinvertebrate metrics to a core set of metrics for further analysis. The multiple linear regression modelling was used to partition the

amount of variation (of the core set of metrics) attributed to the stressors and environmental covariates.

5.3. Methods

5.3.1. Collation of a macroinvertebrate-stressor dataset

A spatial dataset was compiled of sites that spanned a wide gradient of land use (different land uses) and hence also included wide gradients of the two major stressors, nutrients and fine sediment. Collation of existing national and research datasets is described in Sections 2.1.1 and 2.1.2, respectively. However, only subsets of these large datasets were used for linking metrics to stressors. Details on the data collation process including data selection are described in Appendix 4 and a summary is provided here.

Matching the national macroinvertebrate dataset with-stressor data

The national macroinvertebrate dataset was spatially and temporally matched with stressor data retrieved from three separate datasets describing water quality, deposited and suspended sediment, and periphyton. Matching of sites was done in R using various identifiers including site name, regional council site ID, LAWA ID and NZReach ID. Manual checks were also performed to avoid mismatching.

Due to potentially significant annual variation of sediment, nutrient and periphyton conditions at a single site, we calculated a median value from all available stressor data collected within the same month and the 12 months prior to macroinvertebrate sampling. However, monthly (or more) observations were not always available. In particular for deposited sediment measures, often only a single observation was available and hence used as representing the 12-month time period.

Merge with research data

A subset of the research dataset that was also used for stressor-specific metric development was merged with the national macroinvertebrate-stressor dataset. (More details on the research data selection process and rational are described in Section 3.5.2.) The merged dataset was checked for outliers, the influence of sampling methods and data source (national or research dataset) using protocols proposed by Zuur et al. (2010).

Stressor variables

We selected variables that were the most common across the dataset in order to maximise sample size (Table 14). Increased nutrients affect macroinvertebrate communities mainly via algal proliferation in clear streams, therefore we selected chlorophyll-*a* as a measure of stressor intensity. However, we also included DIN and DRP as measures of nitrogen and phosphorus concentrations, respectively, to explore

whether effects can also be attributed to nutrient concentrations directly. As a measure of deposited sediment we selected sediment cover which was visually assessed instream as the percentage of the streambed covered by fine sediment (inorganic particles with diameter < 2 mm). Turbidity was the most common measure of suspended fine sediment but not as common as the other measures. Hence, we decided to exclude turbidity as a predictor because it would have significantly reduced sample size.

Table 14. Mean and range in stressor variables in the collated dataset.

Stressor variable	Description	mean	min	max
CHLA	Periphyton biomass measured as chlorophyll-*a* (mg/m^2)	33.43	0.00	374.00
DRP	Dissolved reactive phosphorus (mg/l)	0.02	0.00	0.40
DIN	Dissolved inorganic nitrogen (mg/l)	0.44	0.01	11.61
instreamVis	Fine sediment (< 2 mm) cover of the streambed from an instream visual assessment (%)	18.68	0.00	100.00

Catchment land use, water abstraction and environmental descriptors

Information on catchment land use (4 variables), surface water abstraction pressure (maxrateToQ50) and 14 relevant environmental variables were retrieved from existing databases via NZReach ID for each macroinvertebrate sampling site (Table 15). Environmental variables describe flow, temperature, geology, catchment morphology, position in the stream network, shading and elevation, and have been previously shown to influence macroinvertebrate communities (e.g. Clapcott et al. 2012; Wagenhoff et al. 2017).

Table 15. Mean values and range in catchment land use and environmental descriptors in the collated dataset, along with their data source and description. Three flow statistics were downloaded on 23 August 2016 from the MfE website: (https://data.mfe.govt.nz/table/2536-natural-river-flow-statistics-predicted-for-all-river-reaches/) and surface water abstraction pressure was downloaded on 16 June 2017 from the MfE website (https://data.mfe.govt.nz/table/3614-accumulated-freshwater-takes-201314/). LCDB3 = Land Cover Data Base 3 (https://lris.scinfo.org.nz/layer/304-lcdb-v30-deprecated/), REC = River Environment Classification database (Snelder et al. 2004), FENZ = Freshwater Ecosystems New Zealand database (Leathwick et al. 2010).

Predictor variables	Source	Description	Mean	Min	Max
Land use					
T1NativeVeg	LCDB3	Percentage catchment land cover in native vegetation including forest and scrubland	35.81	0.00	100.00
T1Urban	LCDB3	Percentage catchment land cover in urban land uses	3.08	0.00	100.00
T1ExoticVeg	LCDB3	Percentage catchment land cover in exotic vegetation including forest and scrubland	11.48	0.00	100.00
T2PastoralHeavy	LCDB3	Percentage catchment land cover in heavy pastoral land uses including exotic grassland, short rotation cropland, orchards and vineyards	40.53	0.00	100.00
maxrateToQ50	MfE website	Estimated impact of upstream consents on the modelled median flow of a particular reach calculated by upstream total consented takes divided by median flow	0.08	0.00	20.03
Environmental					
DSDist2Coast	FENZ	Distance to coast	72.61	0.08	432.84
Elevation	REC	Altitude of the stream segment	142.73	0.00	1142.59
FRE3	MfE website	Annual frequency of flood events > 3x median annual flow	14.23	1.81	37.35
ORDER	REC	Stream order	3.69	1.00	8.00
SegFlowStability	FENZ	Ratio of mean annual low flow/ mean annual mean flow	0.17	0.00	0.54
SegJanAirTemp	FENZ	Summer air temperature for a segment	16.86	10.90	19.70
SegMinTNor	FENZ	Seasonal air temperature range for a segment	0.54	-4.14	2.84
SegRipShade	FENZ	Riparian shade for a segment	0.28	0.00	0.80
SpecMALF	MfE website	Specific mean annual low flow (= mean flow / catchment area)	0.01	0.00	0.06
SpecMeanFlow	MfE website	Specific mean flow (= mean flow / catchment area)	0.03	0.01	0.23
USSlope	FENZ	Average slope in the catchment	12.82	0.04	32.41
USCalcium	FENZ	Average calcium concentration of underlying rocks	1.53	0.39	3.99
USHardness	FENZ	Average hardness of underlying rocks	3.01	0.78	5.00
USPhosphorus	FENZ	Average phosphorus concentration of underlying rocks	2.34	0.78	5.00

Sample size and spatial spread

The collated dataset contained 8,774 macroinvertebrate samples (70% national and 30% research data) from a total of 1,656 sites where at least one stressor variable was also measured (left panel in Figure 15). This dataset was used for analyses linking macroinvertebrate metrics to catchment-scale descriptors of human land use. For analyses which linked macroinvertebrate metrics to reach-scale stressors, a subset of the data was used. The subset contained 510 macroinvertebrate samples (20% national and 80% research data) from a total of 257 sites where all focal stressor variables (i.e. chlorophyll-*a*, DIN, DRP, sediment cover) were measured (right panel in Figure 15).

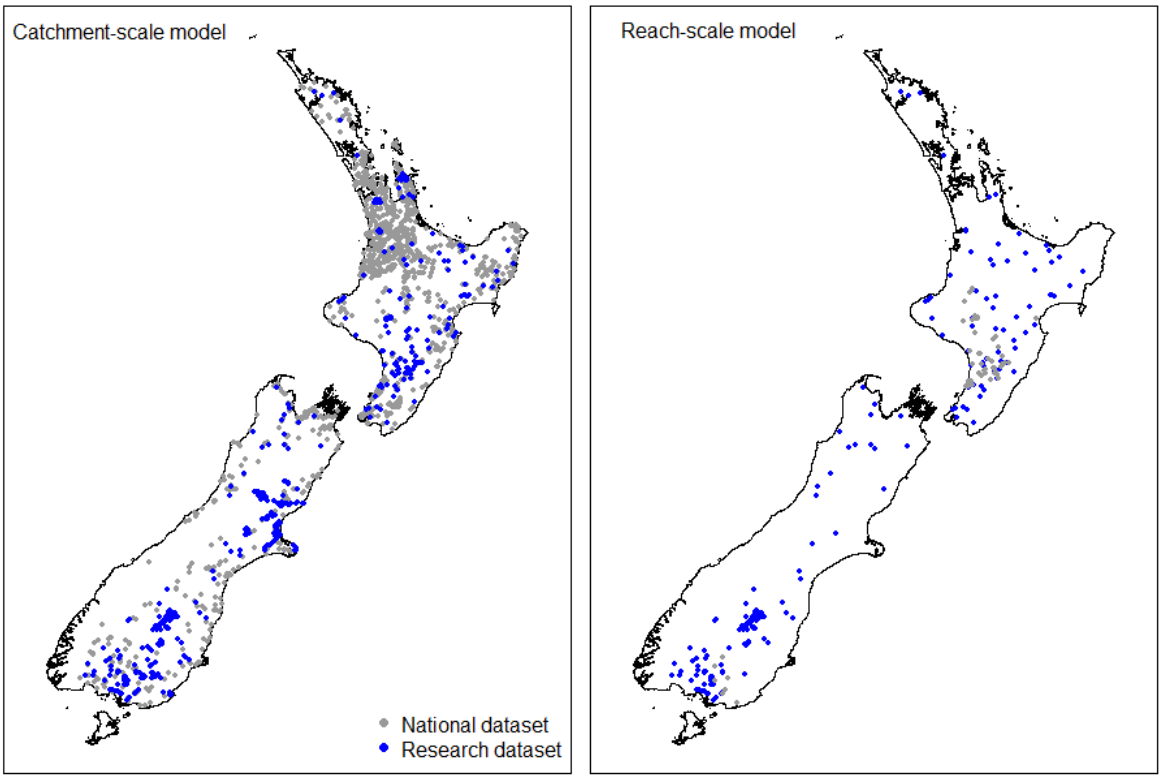


Figure 15. Spatial spread of the data used for statistical analyses to link macroinvertebrate metrics to catchment-scale pressures (land-use cover) (left) or stressor variables assessed at the reach scale (right).

5.3.2. Overview of analytical process

A flow diagram illustrates the steps involved in our analytical process (Figure 16). We first performed a gradient forest (GF) analysis which builds a random forest (RF) model for each metric. Reduction of the initial candidate set of metrics was achieved by selecting only those that were among the best 16 metrics based on maximum cumulative importance retrieved from the GF output for each of the land-use gradients

(catchment-scale models) and also for each deposited fine sediment and chlorophyll-a stressor gradients (reach-scale models) as well as flow allocation pressure. This led to a total of 26 and 28 candidate metrics for the catchment-scale and reach-scale analysis, respectively (Figure 16). Secondly, these metrics were investigated for their response shape to the primary land use (i.e. native vegetation or pastoral cover) or stressor gradients. In particular, concordance with ecological theory, overall effect size and gradual response shape were determinants for inclusion of suitable metrics. We excluded highly correlated metrics as this point. This resulted in 14 candidate metrics from the catchment-scale analysis and 18 from the reach-scale analysis. For this smaller candidate set of metrics, multiple linear regression models were built to statistically confirm the link between metrics and stressors. Selection criteria included linear model R^2, effect size, and the amount of independent variation partitioned to the stressors. The retained metrics can be considered the core set for development of an ecosystem health assessment framework (Section 6).

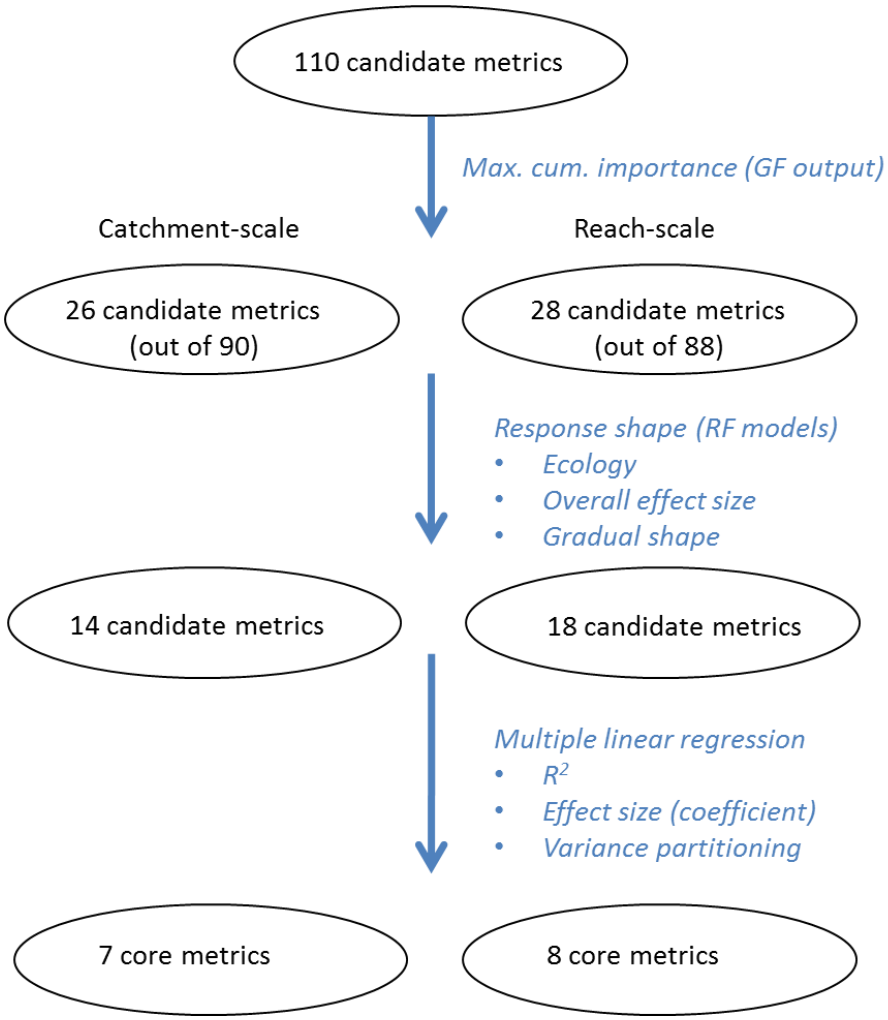


Figure 16. Flow diagram illustrating core metric selection from a large amount of candidate metrics considered for inclusion in a stream ecosystem health assessment framework.

5.3.3. Gradient forest analysis

Gradient forest analysis builds a random forest (RF) model for each metric and was chosen as an exploratory approach to link macroinvertebrate metrics to stressors. RF models (1) automatically take into account interactions between predictors, (2) can describe complex response shapes visualised in partial dependence plots that show the fitted function of a metric to a predictor when all other predictors are held at their mean values, and (3) estimate the relative importance of the predictors. Gradient forest (GF) analysis is implemented in the statistical programme R with specialised functions provided in two R packages. GF analysis is not only an efficient way to build many RF models at once, but its RF functions also use an improved method for calculating variable importance when correlated predictors are present. Furthermore, GF analysis provides a turnover function across each predictor gradient visualising cumulative importance for each metric. The larger the maximum cumulative importance, the larger the importance of the predictor for overall change in this metric. A short description of the computational method of the GF analysis is provided in Section 3.5.1 and a full description can be found in Ellis et al. (2012).

Reach-scale and catchment-scale models

We conducted GF analysis twice depending on whether catchment land-use or reach-scale stressors (nutrients, periphyton and sediment) were included as predictors. The set of environmental descriptors and the water abstraction pressure estimate (maxRateToQ50) were included as predictors in both analyses. Metrics were transformed to aid visualisation of response shapes using the Yeo-Johnson power transformation to approximate normal distribution. The Yeo-Johnson transformation is similar to the Box-Cox model but can accommodate predictors with zero and/or negative values. After transformation, all metrics were centred and scaled (by subtracting the overall mean from each observation and dividing the result by the overall standard deviation) to make effect sizes directly comparable. The predictors do not need to be transformed for RF modelling. Overall, we considered 110 macroinvertebrate metrics including (1) 31 existing stream health metrics including diversity metrics but excluding ASPM (which is a multi-metric index) as well as semi-quantitative and soft-bottom MCI metrics (Table 1 described in Sections 2.2 and 2.3); (2) 59 trait modalities, (Table 4 in Section 2.4), and (3) 20 newly-developed stressor-specific metrics (10 per stressor, described in 3.4.2). The catchment-scale (land use) analysis included RF models for 90 candidate metrics, i.e. all 110 metrics excluding the 20 stressor-specific metrics. The reach-scale (stressor) analysis included 88 candidate metrics, i.e. the 110 metrics excluding diversity metrics and those which were strongly correlated, for example, MCI_hb but not MCI_hb2.

5.3.4. Multiple linear regression analysis

Linear regression analysis was used to quantify the relationship between stressors and metrics. It assumes linear relationships between predictor and response

variables. We developed multiple linear regression models for selected metrics from the GF analysis in order to partition the variance among the stressors (catchment-scale or reach-scale) as well as covariates. The aim of this analysis was to statistically confirm the suitability of macroinvertebrate metrics as either indicators of general stream health or stressor-specific metrics in an EH assessment framework.

Reach-scale models

Multiple linear regression models were used to determine the relationship between each macroinvertebrate metric and the stressors and environmental predictor variables. Macroinvertebrate metrics were those that responded best to the three focal stressors (CHLA, instreamVis, maxrateToQ50), but excluding highly correlated metrics (e.g. including Chl_MCI_like but excluding Chl_MCI_like_log) (Table 16). Initial data exploration was conducted following the protocol proposed by Zuur et al. (2010) to check for outliers and the normality of variable distributions. All response variables (metrics) and predictor variables (stressors) were transformed to improve normality using the Yeo-Johnson transformation. After transformation, all variables were centred and scaled before the analyses (in the same way as for GF analysis, see Section 5.3.3), to allow direct comparison of regression coefficients and inference about relative effects sizes among stressors and metrics. Correlations among predictor variables and among metrics were visualised using correlograms (not shown). Collinearity among predictor variables was checked using a variance inflation factor (VIF, Zuur et al. 2010). Variables with the highest VIF (SpecMeanF and ELEVATION) were sequentially dropped so that all VIFs were < 3. Linear models were fitted with all remaining predictor variables as fixed effects and final predictor variables were selected using a backwards procedure based on the generalised Schwarz's Bayesian information criterion (BIC). Final model diagnostics were checked by plotting residuals versus fitted and normal Q-Q plots of the residuals.

Hierarchical partitioning of R^2 values for each metric was used to determine the proportion of variance explained independently by each variable (Chevan & Sutherland 1991; Mac Nally 2000). This method allows identification of variables whose independent correlation with the dependent variable is strong, in contrast to variables that have little independent effect but have a high correlation with the dependent variable resulting from joint correlation with other independent variables.

Table 16. Mean and range of response variables (macroinvertebrate metric or trait) explored in reach-scale multiple linear regressions. N = 510 samples. Response variables had previously been selected using the GF analysis. For a description of metrics see Table 1, for traits see Table 4 and for 'chl' and 'sed' metrics see Section 3.5.

Metric	Mean	Min	Max	Trait	Mean	Min	Max
chl_MCI_like	102.52	20.00	186.67	1b - SIZE2	0.40	0.05	0.71
chl_pct_richness_decreaser	23.99	0.00	66.67	3b - UNIV	0.55	0.07	0.96
chl_pct_richness_increaser	18.25	0.00	50.00	3c - PLURIV	0.42	0.01	0.93
chl_richness_decreaser	4.04	0.00	13.00	6b - HERMA	0.04	0.00	0.32
chl_richness_increaser	2.85	0.00	6.00	6c - TWO	0.88	0.35	1.00
EPTrich*	6.11	0.00	21.00	7b - SUBMERGED	0.53	0.20	0.99
MCI_hb	104.62	51.11	172.86	8a - EGGFREE	0.59	0.06	0.93
pEPTabund	39.17	0.00	94.44	10b - CRAWLER	0.77	0.58	0.99
pEPTrich*	34.88	0.00	80.00	11b - LOWFLEX	0.34	0.00	0.75
sed_MCI_like	133.17	40.00	183.33	13b - SCRAPER	0.69	0.39	0.94
sed_pct_richness_decreaser	46.72	0.00	100.00	13d - FILTERFEED	0.08	0.00	0.29
sed_pct_richness_increaser	20.46	0.00	77.78				
sed_richness_decreaser	7.55	0.00	19.00				
sed_richness_increaser	3.19	0.00	9.00				

Catchment-scale models

Catchment-scale models were used to quantify the relationship between metrics in Table 17 and land use. They were fitted in the same way as described above for the reach-scale models. Metrics were chosen from the GF output that showed those metrics most associated with each land-use pressure gradient (T1NativeVeg, T2PastoralHeavy, T1ExoticVeg, T1Urban, maxRate ToQ50), but excluding highly correlated variables (e.g. including MCI_hb but not MCI_hb2).

Table 17. Mean values and range of responses (macroinvertebrate metric or trait variables) explored in catchment-scale multiple linear regressions. N = 8,774 samples. Response variables were selected using the GF output. * = excluding Hydroptilidae. For a description of metrics see Table 1 and for traits see Table 4.

Metric	Mean	Min	Max	Trait	Mean	Min	Max
AMDI	48.07	0.00	88.53	1b - SIZE2	0.35	0.00	1.00
EPTrich*	8.18	0.00	29.00	1e - SIZE5	0.02	0.00	0.38
LIFENZ	7.57	4.00	9.21	2a - DESC1	0.34	0.00	1.00
MCI_hb	105.63	30.00	180.00	2b - DESC2	0.57	0.00	1.00
pEPTabund*	41.61	0.00	100.00	2d - DESC4	0.03	0.00	0.50
pEPTrich*	39.63	0.00	100.00	3b - UNIV	0.63	0.00	1.00
QMCI_hb	4.94	1.06	9.27	3c - PLURIV	0.34	0.00	1.00
QUCI	0.74	-0.64	1.85	4a - CPI1	0.75	0.00	1.00
Simpsons	0.69	0.00	0.95	4c - CPI2	0.25	0.00	1.00
totRich	19.99	1.00	51.00	5a - LDA1	0.02	0.00	0.14
UCI	15.00	-10.28	37.91	5d - LDA4	0.27	0.00	1.00
				6c - TWO	0.85	0.26	1.00
				7a - SURFACE	0.34	0.00	1.00
				7b - SUBMERGED	0.56	0.00	1.00
				8c - EGGPROTECTED	0.15	0.00	1.00
				11b - LOWFLEX	0.39	0.00	0.90
				12a - STREAMLINED	0.10	0.00	0.38
				13c - DEPOSIT	0.05	0.00	0.35
				14c - GENERALIST	0.71	0.00	1.00
				15b - GILL	0.48	0.00	1.00
				15c - PLASTRON	0.06	0.00	0.52
				16a - ADUANDLAR	0.33	0.00	1.00
				16b - ADUORLAR	0.25	0.00	1.00

5.4. Results

5.4.1. Gradient forest

Catchment-scale models

Gradient forest analytical output: Overall, the land use stressors of T1NativeVeg, T2PastoralHeavy, T1ExoticVeg, maxrateToQ50 and T1Urban were ranked 6, 8, 16, 17 and 18, respectively, among a total of 19 predictors in the GF models for 90 candidate metrics. These ranks reflect in part the proportion of sites subject to each land use stressor, i.e. fewer sites were subject to urban pressures compared to pastoral pressure at the national scale. The metrics that were most associated with these land use stressors were identified by the maximum cumulative importance scores from the GF turnover function (Table 18; see also Appendix 5). The larger the maximum cumulative importance, the larger the importance of the land use stressor for overall change in this metric. For example, the MCI_hb showed the most important

change across T1NativeVeg and the 3rd-most important change across T2PastoralHeavy.

Table 18. List of the 16 highest-ranked macroinvertebrate metrics including traits for each the percentage cover of native vegetation (T1NativeVeg) and heavy pastoral land use (T2PastoralHeavy) according to their maximum cumulative importance in a gradient forest model. Metrics ordered by response to T1NativeVeg. *excluding Hydoptilidae.

	T1NativeVeg		T2PastoralHeavy	
Metric	**Maximum cumulative importance**	**Rank**	**Maximum cumulative importance**	**Rank**
MCI_hb	0.076	1	0.059	3
11b - LOWFLEX	0.068	2		
EPTrich*	0.068	3		
2d - DESC4	0.068	4		
16b - ADUORLAR	0.067	5	0.054	10
EPTrich	0.064	6		
UCI	0.061	7	0.054	9
pEPTrich*	0.06	8		
pEPTrich	0.06	9		
QMCI_hb	0.059	10		
QUCI	0.058	11	0.053	11
MCI_hb2	0.056	12	0.056	6
3c - PLURI	0.056	13	0.048	16
2a - DESC1	0.055	14	0.06	2
3b - UNIV	0.053	15		
AMDI	0.048	16		
8c - EGGPROTECTED			0.065	1
16a - ADUANDLAR			0.058	4
4a - CPI1			0.057	5
7b - SUBMERGED			0.055	7
4c - CPI2			0.055	8
7a - SURFACE			0.052	12
5d - LDA4			0.052	13
QMCI_hb2			0.05	14
2b - DESC2			0.049	15

Random forest model output: Generally, the higher-ranked metrics, such as the 16 best presented in Table 18, also had higher model R^2 values than lower-ranked metrics (Table A5.1, Appendix 5). R^2 values of the random forest models built during GF analysis ranged from 0.14 to 0.81 with a median of 0.49. Within each random forest model the relative importance of the predictors for each metric or trait is calculated. Native vegetation was ranked in the top 5 predictors for 61% of metrics

whereas heavy pasture cover ranked in the top five predictors for only 17% of metrics; Urban land use was not in the top 5 predictors for any metrics and maxrateToQ50 was the top predictor for one metric (Simpsons diversity) (Table A5.1, Appendix 5).

Response shapes and effect sizes: All macroinvertebrate metrics showed a relatively similar gradual increase across the entire gradient of native vegetation cover in the catchment (Figure 17A). Among those, EPT richness with or without Hydroptilidae resembled each other the most and had the largest relative effect size (Figure 17A). Response shape and effect size of the three traits, low body flexibility, maximum no. of descendants > 3000, and adult or larval aquatic stage were also very similar compared to each other (Figure 17B) and also compared to the metrics. All macroinvertebrate metrics also showed a very similar gradual but negative response to heavy pastoral land cover (Figure 18A). However, overall effect size was slightly lower than that across the native vegetation cover gradient. A larger proportion of traits responded to the pastoral land cover gradient (Figure 18B) compared to the native vegetation cover gradient (Figure 17B). In two cases, two modalities of the same trait responded in opposite directions reflecting that the proportion of one modality decreases at the expense of the other increasing (Figure 18B).

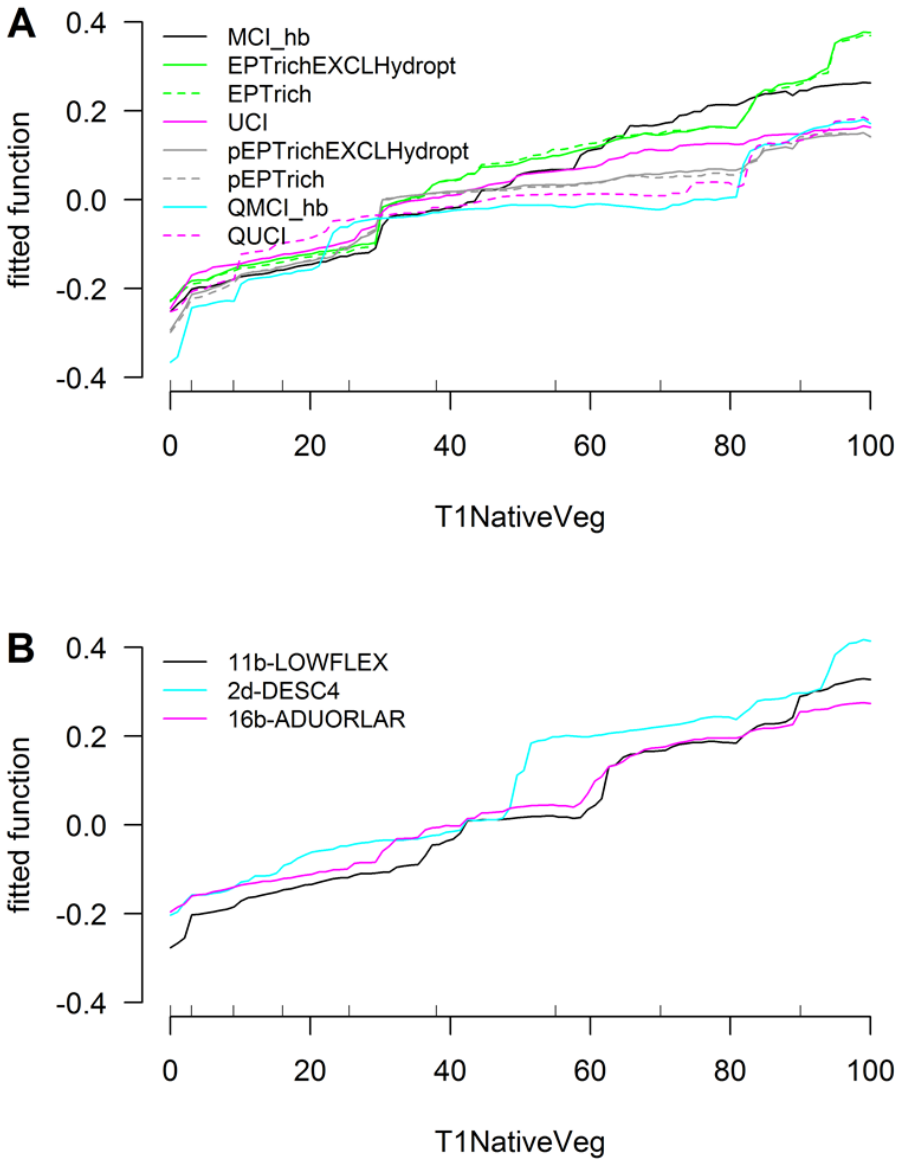


Figure 17. Partial dependence plots of metrics (A) and traits (B) that ranked among the best metrics according to maximum cumulative importance (presented in Table 18). Solid and dashed lines of the same colour indicates two metrics that are very similar.

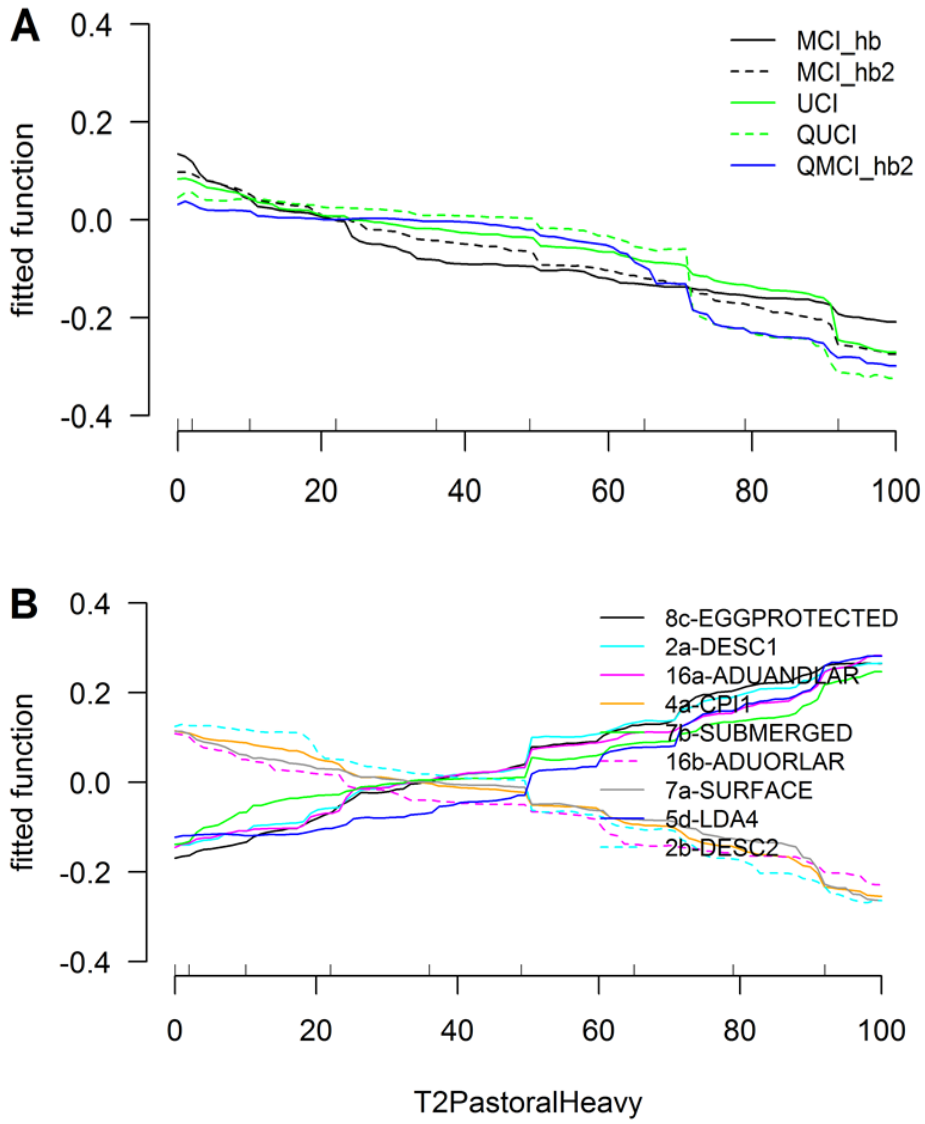


Figure 18. Partial dependence plots of metrics (A) and traits (B) that ranked among the best metrics according to maximum cumulative importance (presented in Table 18). Dashed lines and same colour indicates in A that the two metrics are very similar, and in B that the trait modalities are from the same trait.

Reach-scale models

The R^2 values of the individual random forest models built during GF analysis ranged from 0.03 to 0.67 with a median of 0.52. Within each random forest model the relative importance of the predictors for each metric or trait was calculated. Sediment cover was ranked in the top 5 predictors for 11 out of 25 metric or traits, whereas chlorophyll-*a* ranked in the top five predictors for 17 out of 25; DIN was in the top 5 predictors for 5 out of 25 metrics or traits and DRP was in the top 5 predictor for 8 out of 25 (Table A5.3 and Table A5.4, Appendix 5). These results confirmed our

assumption that sediment cover was an important descriptor of the intensity of sediment effects and that chlorophyll-*a* was a better descriptor of the intensity of enrichment effects than the nutrient concentrations themselves.

The GF analysis was applied to select the best metrics among a large set of metrics to be further investigated with multiple linear regression models. The first selection step involved gradient forest analytical output. Overall, the proximate stressors of DRP, chlorophyll-*a* (CHLA), DIN and deposited sediment (instreamVis) were ranked 2, 3, 10 and 12, respectively, among a total of 19 predictors in the GF models for 89 candidate metrics. The water abstraction estimate ranked 4th in the reach-scale models. The highest-ranking metrics across the chlorophyll-*a* and deposited sediment gradients were the stressor-specific metrics developed in Section 3 (Table 19). Ranked first for the % sediment cover (instreamVis) gradient was the number of sediment tolerant taxa (sed_richness_increaser) and there were five further sediment-specific metrics but only a single nutrient-specific metric (chl_richness_decreaser = number of enrichment sensitive taxa) among the 16 best metrics (Table 19). Similarly, ranked first for the chlorophyll-*a* (CHLA) gradient was the number of enrichment tolerant taxa (chl_richness_increaser).There were four further nutrient-specific metrics but only a single sediment-specific metric (sed_pct_richness_decreaser = percentage of sediment-sensitive taxa) among the 16 best metrics (Table 19). Generally, there was little overlap of the best metrics among the two focal proximate stressor gradients (Table 19). Three general stream health macroinvertebrate metrics (MCI and EPT types) also ranked among the 16 best for either only CHLA (MCI_hb, rank 11) or both stressor gradients (pEPTabund with and without Hydroptilidae) (Table 19). Finally, 13 traits also were among the best 16 metrics and as for stressor-specific metrics there also was little overlap of the best metrics among the two stressor gradients (Table 19). The metric ranks for all four stressor gradients are given in Appendix 5 (Figure A5.3 and Figure A5.4).

Table 19. List of the 16 highest-ranked metrics including traits for each focal proximate stressor attribute, % sediment cover (instreamVis) and chlorophyll-*a* (CHLA) according to their maximum cumulative importance. *excluding Hydoptilidae.

	instreamVis		CHLA	
Metric	**Maximum cumulative importance**	**Rank**	**Maximum cumulative importance**	**Rank**
sed_richness_increaser	0.052	1		
sed_pct_richness_increaser	0.049	2		
pEPTabund	0.048	3	0.081	8
chl_richness_decreaser	0.046	4		
7b - SUBMERGED	0.045	5		
pEPTabund*	0.045	6	0.1	2
sed_pct_richness_decreaser	0.044	7	0.07	10
10b - CRAWLER	0.042	8		
2c - DESC4	0.04	9		
5a - LDA1	0.039	10	0.099	4
sed_MCI_like_log	0.039	11		
sed_MCI_like	0.038	12		
16a - ADUANDLAR	0.036	13		
1a - SIZE1	0.036	14		
3c - PLURI	0.035	15		
sed_richness_decreaser	0.035	16		
chl_richness_increaser			0.102	1
chl_MCI_like			0.099	3
chl_MCI_like_log			0.097	5
11b - LOWFLEX			0.097	6
6b - HERMA			0.087	7
chl_pct_richness_decreaser			0.073	9
MCI_hb			0.07	11
chl_pct_richness_increaser			0.063	12
16b - ADUORLAR			0.054	13
3b - UNIV			0.053	14
9a - DISSLOW			0.052	15
6c - TWO			0.05	16

The second selection step involved visual investigation of response shapes and effect sizes from the partial dependence plots of the random forest models. The response shapes of the fitted functions of the 16 most important metrics including traits were plotted for each respective stressor variable. All sediment-specific metrics responded as expected. The metrics based on tolerant taxa (increasers) increased while those based on sensitive taxa (decreasers) decreased with increasing sediment cover (Figure 17A), and the 'Sediment MCI' incorporating tolerance values of increasers and decreasers (sed_MCI_like, sed_MCI_like_log) decreased across the sediment

gradient. The overall effect size across % sediment cover was similar among sediment-specific and general stream health metrics, however response shapes somewhat differed among metrics. For example, the number of sediment-sensitive taxa (sed_richness_decreaser) and the 'Sediment MCI' (sed_MCI_like_log) showed a relatively gradual decrease while %EPT abundance (pEPTabund) showed a more sudden change at about 60% sediment cover (Figure 19A). EPT richness with or without Hydroptilidae and sed_MCI_like and sed_MCI_like_log resembled each other the most, respectively (Figure 19A). Some traits showed effect sizes similar to the stressor-specific or EPT metrics although other traits showed smaller effect sizes, e.g. Aduandlar (Figure 19B), but response shapes were relatively gradual.

All nutrient-specific metrics also responded as expected across the chlorophyll-*a* gradient. The metrics summarising tolerant taxa (increasers) increased while those summarising sensitive taxa (decreasers) decreased (Figure 20A) and the 'Nutrient MCI' decreased (chl_MCI_like, chl_MCI_like_log) as well. General stream health metrics (MCI_hb, pEPTabund with and without Hydroptilidae) also decreased as expected. Response shapes were relatively similar across the stream health and nutrient-specific metrics, with a relatively steep decrease or increase early on in the full observed stressor gradient and no further change after (Figure 20A). The shapes were also similar among the traits (increase or decrease) but effect sizes varied (Figure 20B).

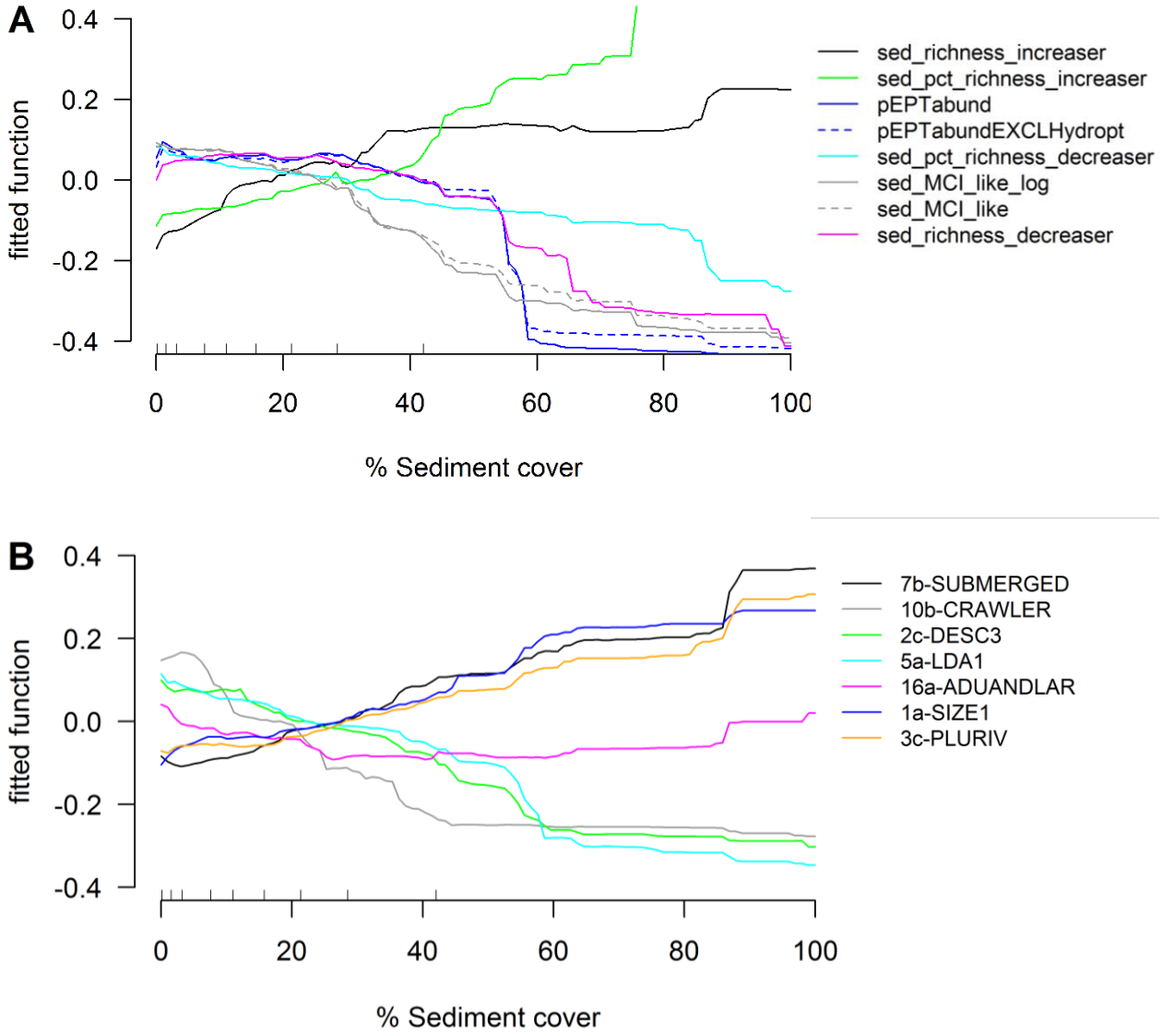


Figure 19. Partial dependence plots of metrics (A) and traits (B) that ranked among the 16 best metrics in response to instream visual assessment of % sediment cover (instreamVis) according to maximum cumulative importance (presented in Table 19). In A, dashed lines and same colour indicates that the two metrics are very similar.

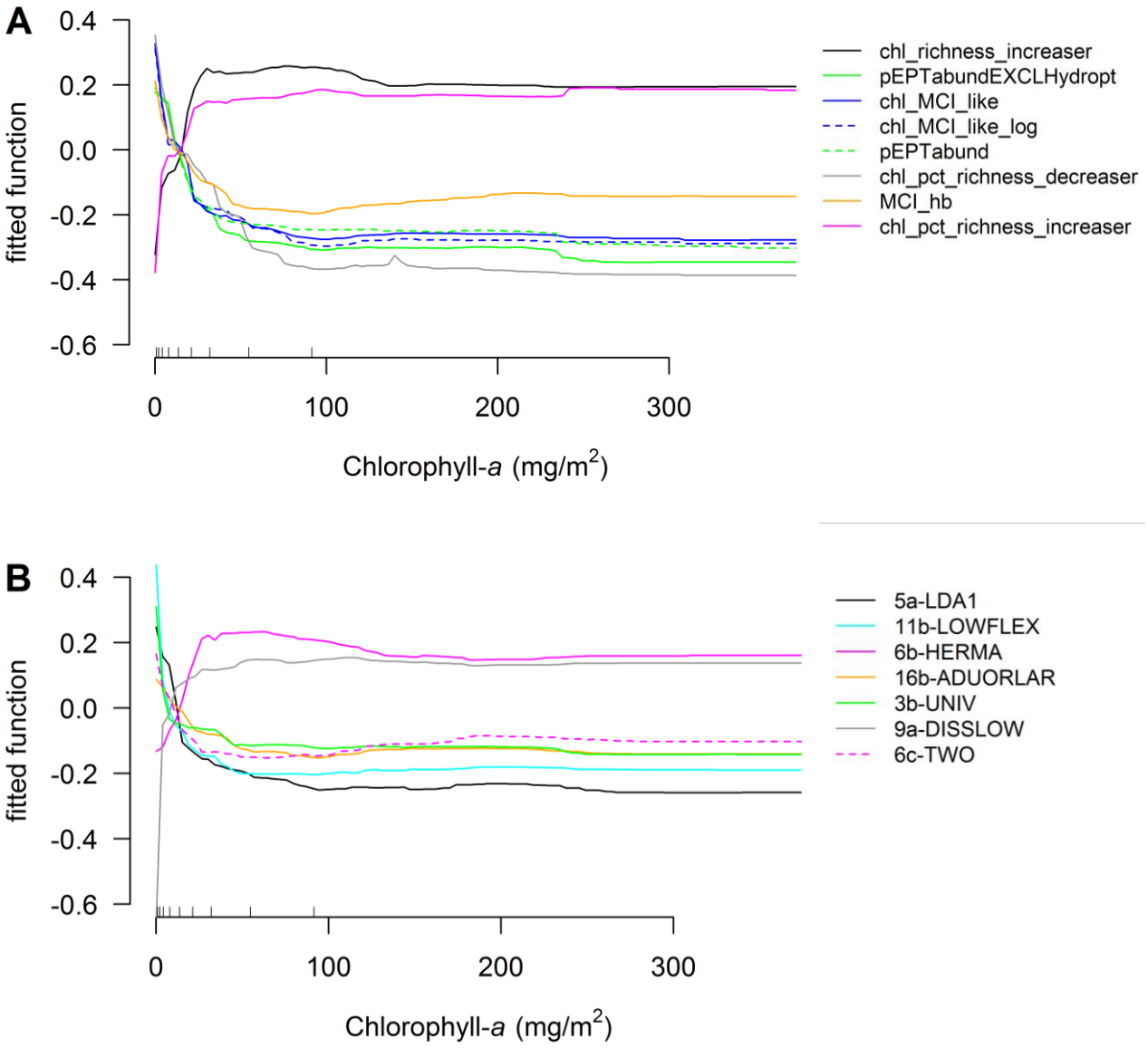


Figure 20. Partial dependence plots of metrics (A) and traits (B) that ranked among the 16 best metrics in response to chlorophyll-*a* (CHLA) according to maximum cumulative importance (presented in Table 19). Dashed lines and same colour indicates in A that the two metrics are very similar, and in B that the trait modalities are from the same trait.

5.4.2. Multiple linear regression

Catchment-scale models

All catchment-scale regression models were highly significant ($P < 0.001$ for all metrics and traits). The amount of variation explained (R^2) ranged between 0.09 and 0.58, with LIFENZ, UCI and MCI_hb and pEPTrich (excl. Hydroptilidae) models having the largest R^2 values (> 0.4) among metrics. Among traits, the models for 4a-CPI1 (one reproductive cycle per individual), 4c-CPI2 (two or more reproductive

cycles per individual) and 16a-Aduandlar (adult and larva aquatic life stages) had the largest R^2 values (> 0.48).

Multiple linear regression models for response variables with $R^2 > 0.4$ were explored further to (1) estimate specific effect sizes of the catchment land-use pressures and (2) partition the independent variance of the predictors. The specific effect sizes of catchment-scale land-use pressures and environmental predictors (i.e. their regression coefficients along with their confidence intervals) are presented separately for metrics (Figures 21 and 22) and traits (Figures 23 and 24). A summary of regression coefficients and confidence intervals for each predictor variable included in all models is shown in Appendix 5.

Metrics

Regarding effect sizes, T1NativeVeg had the largest absolute effect size (mean = 0.32, 95% CI = 0.29–0.33) and was included in three out of the four metric models with $R^2 > 0.4$. T1ExoticVeg (mean = 0.08, 95% CI = 0.06–0.10) and T1Urban (mean = -0.08, 95% CI = -0.10 to -0.06) were included in the four metric models and heavy pasture was only included in two models (Figure 21).

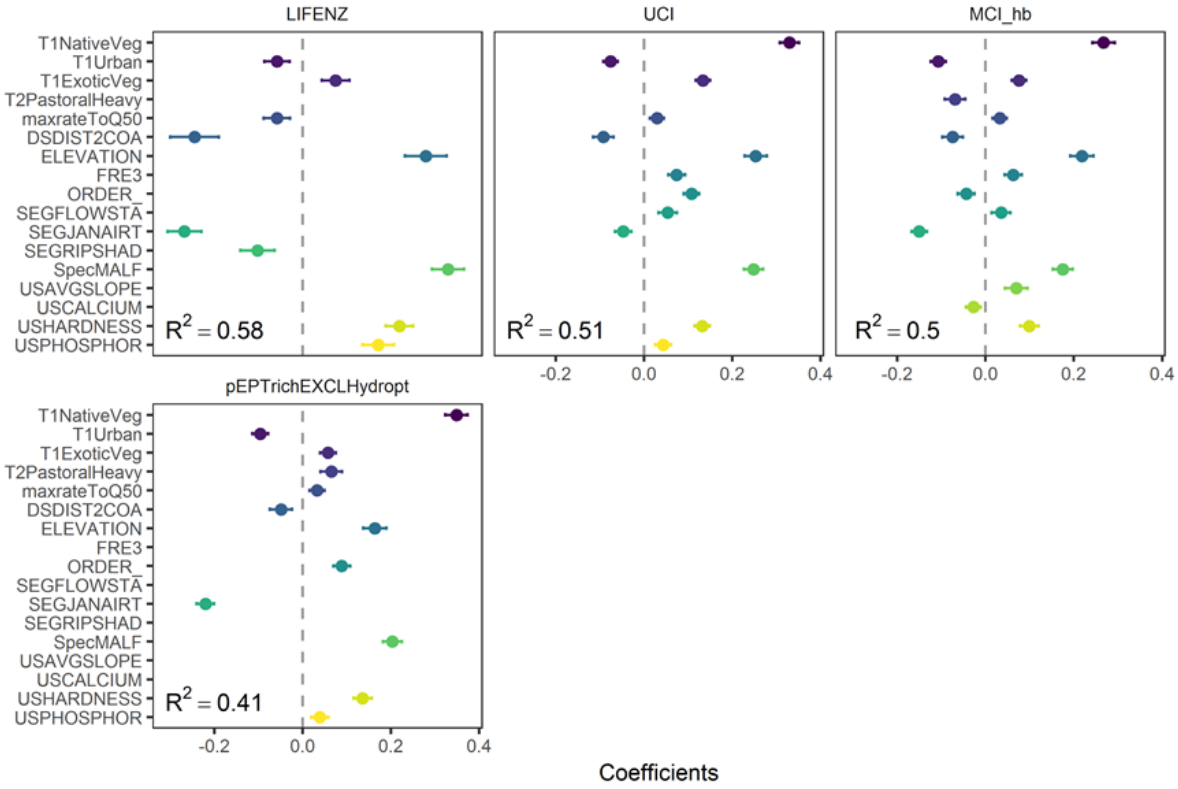


Figure 21. Regression coefficients (± 95% CI) of fixed effects obtained from linear models for each metric with $R^2 > 0.4$. Metrics are ordered by decreasing R^2 values. Predictors on the y-axis are land use stressors (i.e. T1NativeVeg, T1Urban, T1ExoticVeg, T2PastoralHeavy and maxrateToQ50) followed by environmental predictors. Predictor variables without regression coefficients were excluded during model selection.

Hierarchical partitioning of variance showed that overall, T1NativeVeg had the highest relative importance in metric models, explaining on average 25% (± 3% SE) of the total variance (Figure 22). For the LIFENZ metric, land use predictors (i.e. T1NativeVeg, T1Urban, T1ExoticVeg, T2PastoralHeavy and maxrateToQ50) together only accounted for 9% of the total variance explained by the model (R^2 = 0.58, Figure 22), whereas for UCI, MCI_hb and pEPTrich (excluding Hydroptilidae) land use stressors accounted for 38%, 38% and 44% of the total variance explained by these models, respectively (Figure 22).

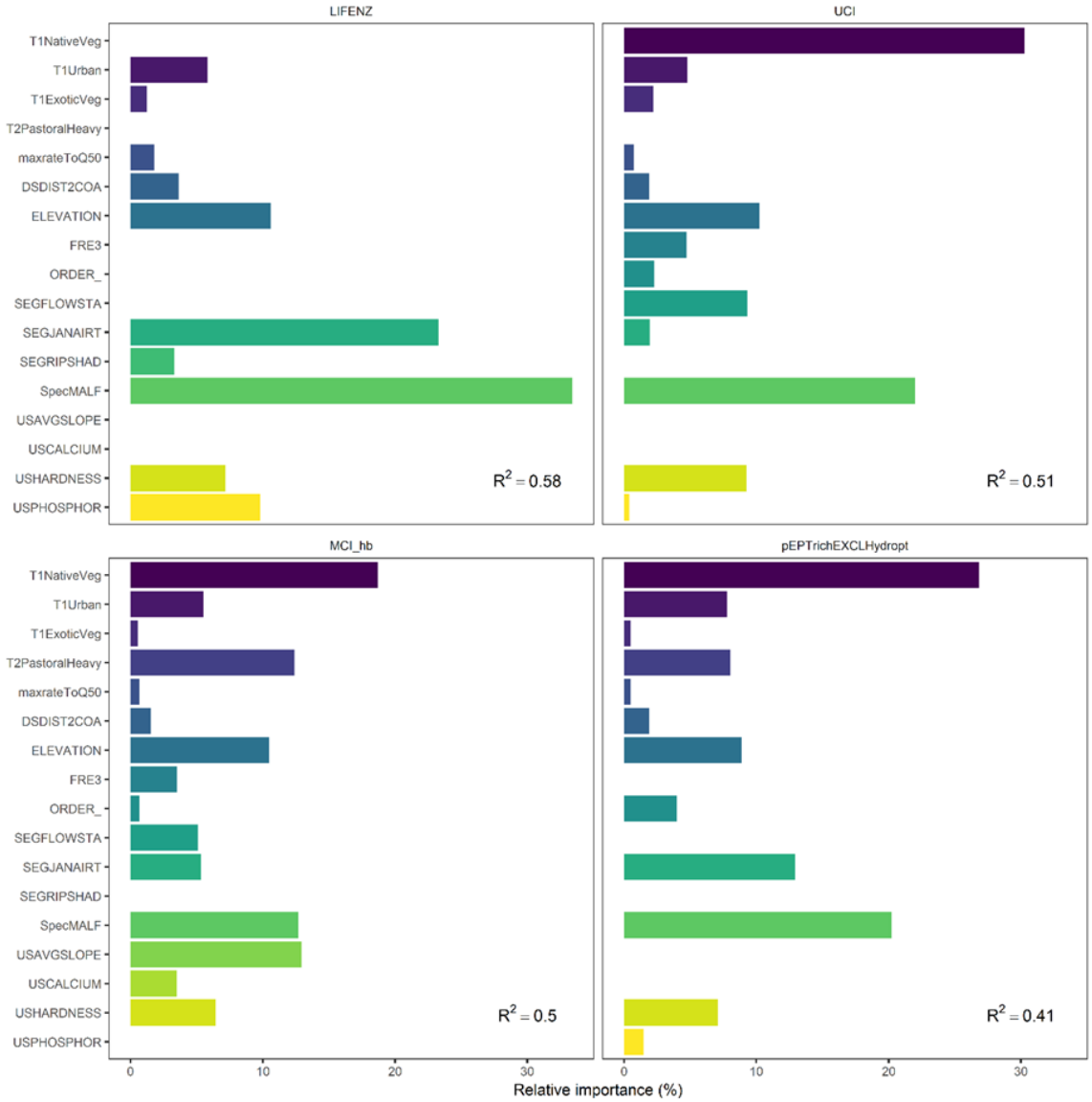


Figure 22. Relative importance of predictor variables from linear models for each metric with $R^2 > 0.4$. Metrics are ordered by decreasing R^2 values and lands use stressors are listed at the top (i.e. T1NativeVeg, T1Urban, T1ExoticVeg, T2PastoralHeavy and maxrateToQ50), followed by environmental predictors. Predictor variables not explaining any variance were excluded during model selection.

Traits

The specific effects of land use and environmental drivers were examined for ten traits where multiple linear regression model $R^2 > 0.4$:

- 2a-DESC1 (≤ 100 descendants per reproductive cycle)
- 4a-CPI1 (1 reproductive cycle per individual)
- 4c-CPI2 (≥ 2 reproductive cycles per individual)
- 5a-LDA1 (≤ 1 day life duration of adults)
- 6c-TWO (male and female reproductive technique)

- 7a-SURFACE (water surface oviposition site)
- 7b-SUBMERGED (submerged oviposition site)
- 8c-EGGPROTECTED (female bears eggs in/on body)
- 16a-ADUANDLAR (Adult and larva aquatic life stages)
- 16b-ADUORLAR (adult or larva aquatic life stages).

T1NativeVeg had the largest absolute effect size (mean = 0.20, 95% CI = 0.18–0.22) and was included in nine out of the ten trait models; followed by T2PastoralHeavy (0.08, 95% CI = 0.05–0.10) and T1ExoticVeg (0.07, 95% CI = 0.06–0.9), which were included in eight and all of ten trait models, respectively (Figure 23).

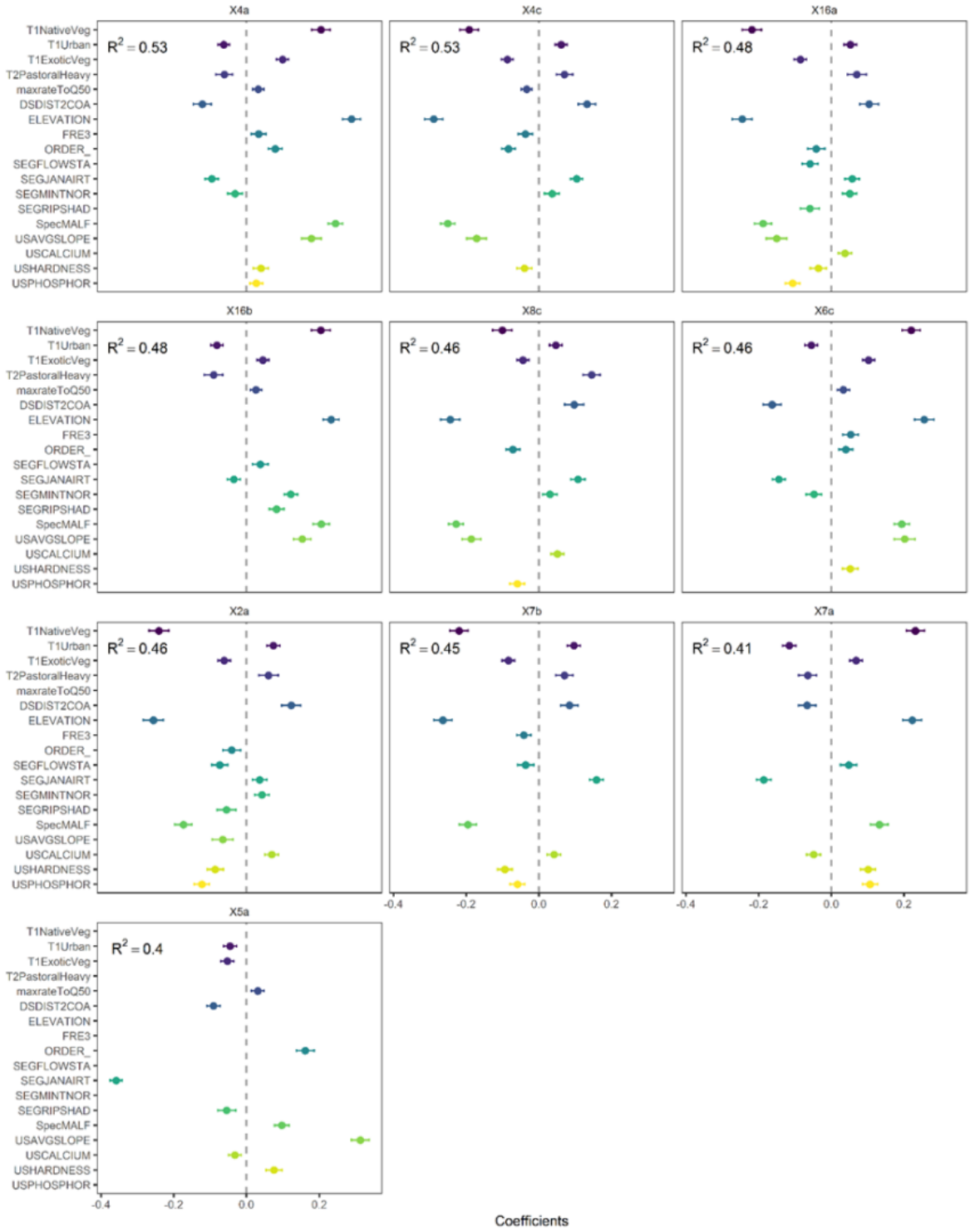


Figure 23. Regression coefficients (± 95% CI) of fixed effects obtained from linear models for each trait with $R^2 > 0.4$. Traits are ordered by decreasing R^2 values (see text for a description of traits). The y-axis shows land use stressors followed by environmental predictors. Predictor variables without regression coefficients were excluded during model selection.

Hierarchical partitioning of variance showed overall, USAVGSLOPE had the highest relative importance for trait models, comprising on average 19% (± 2% SE) of the total variance (R^2) explained by the ten traits. This was followed by T1NativeVeg and SpecMALF, with average relative importance of 17% (± 1 S.E.) and 15% (± 1% S.E.), respectively (Figure 24). Land-use predictors (i.e. T1NativeVeg, T1Urban, T1ExoticVeg, T2PastoralHeavy and maxrateToQ50) together accounted between for 5% and 40% of the total variance explained by these traits models (Figure 24), with the largest proportion for 16b-Aduorlar (40%, R^2 = 0.48) and 7a-Surface (38%, R^2 = 0.41). The relative importance of each predictor variable retained during model selection for the ten traits models is shown in Figure 24.

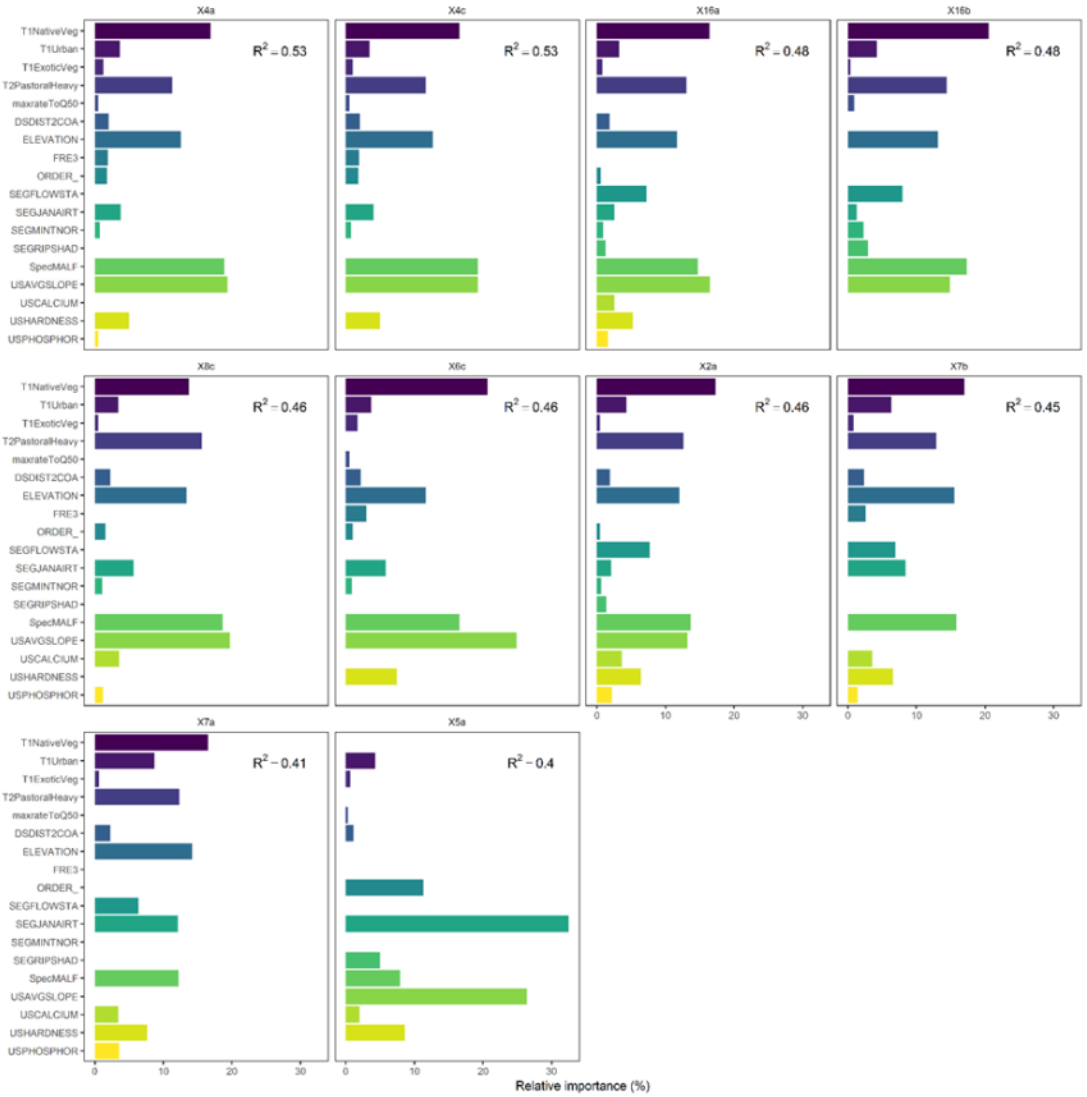


Figure 24. Relative importance of predictor variables from linear models for each metric with R^2 > 0.4. Metrics are ordered by decreasing R^2 values and land-use stressors are listed at the top (i.e. T1NativeVeg, T1Urban, T1ExoticVeg, T2PastoralHeavy and maxrateToQ50), followed by environmental predictors. Predictor variables not explaining any variance were excluded during model selection.

Reach-scale models

All reach-scale regression models were highly significant ($P < 0.001$ for all metric and traits). The amount of variation explained (R^2) by models ranged between 0.09 and 0.51. The metric models with largest R^2 values were sed_pct_richness_decreaser, MCI_hb, sed_richness_increaser, sed_MCI_like and chl_MCI_like. The trait models with the largest R^2 values were 11b-Lowflex (low body flexibility), 7b-Submerged (submerged oviposition site), 3c-Pluriv (purivoltine reproductive cycles) and 6c-Two (male and female reproductive technique). The specific effects of stressors and environmental predictors were examined for a subset of the response variables, having a $R^2 > 0.3$, and are presented below separately for metrics and traits. Summary of all regression coefficients for each predictor variable included in the metric and trait stressor models are shown in Appendix 5.

Metrics

Regarding effect sizes, of the three stressor predictor variables included in the models, chlorophyll-*a* (CHLA) had the largest absolute effect size (mean = 0.21, 95% CI = 0.17 - 0.25) and was included in seven out of the 11 metric models with $R^2 > 0.3$. Instream deposited fine sediment (instreamVis) (mean = 0.17, 95% CI = 0.13–0.20) and allocated flow relative to mean flow (maxrateToQ50) (mean = 0.11, 95% CI = 0.04–0.18) were included in eight and five of the 11 metric models explored, respectively (Figure 25).

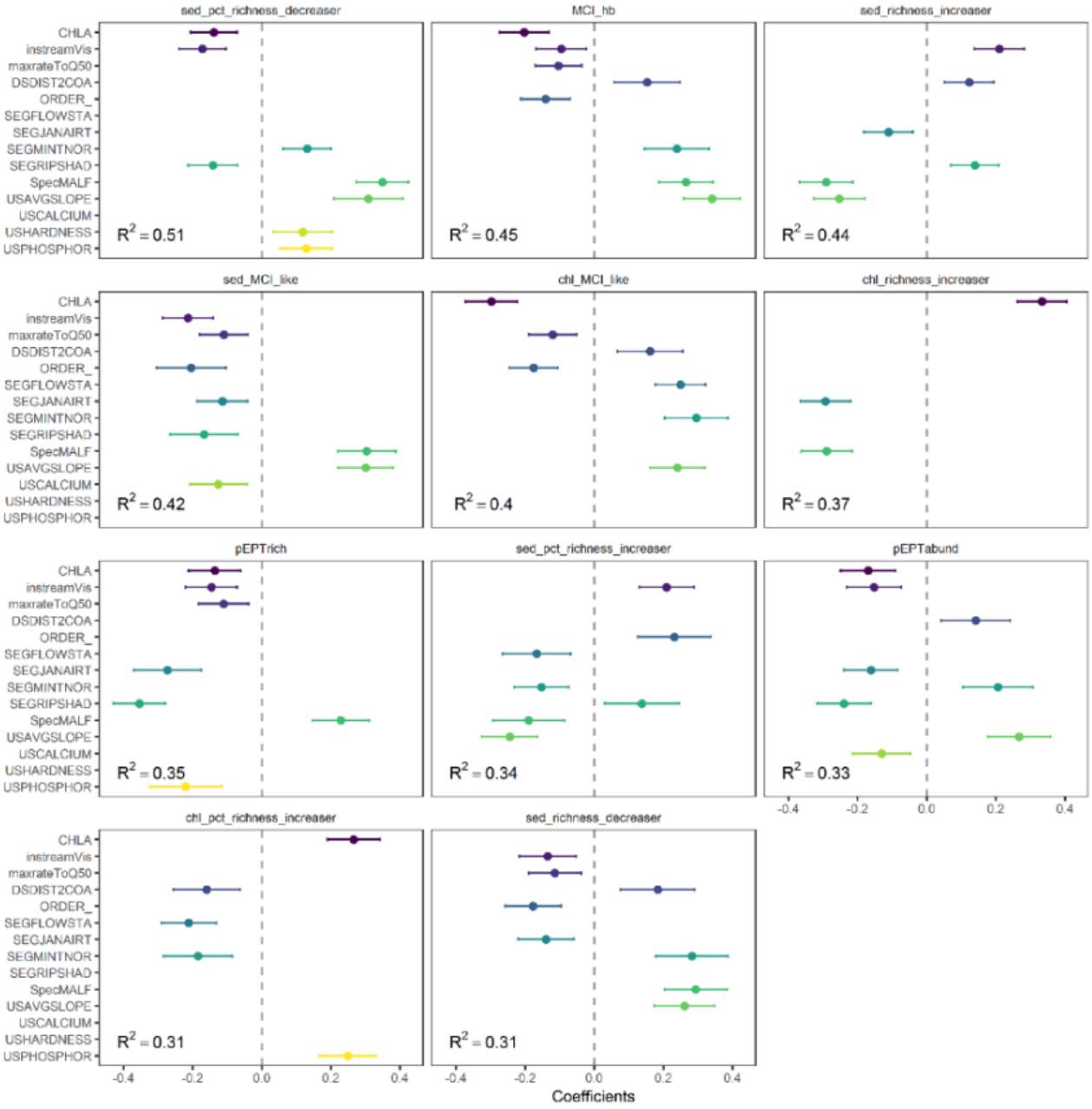


Figure 25. Regression coefficients (± 95% CI) of fixed effects obtained from linear models for each metric with $R^2 > 0.3$. Metrics are ordered by decreasing R^2 values, displayed in each panel. EPT metrics are excluding Hydroptilidae. Predictor variables without regression coefficients were excluded during model selection.

Hierarchical partitioning of variance showed that overall, CHLA was the stressor that had the largest relative importance for metrics models, explaining on average 24% (± 4% SE) of the total variance (R^2) in the data. This was followed by instreamVis and maxrateToQ50, which had an overall average importance of 14% (± 2 SE) and 4% (± 1% SE), respectively (Figure 26). Metrics with greatest independent variance

attributable to CHLA alone were chl_richness_increaser and chl_pct_richness_increaser. While MCI_hb and chl_MCI_like had large independent variance attributed to CHLA, there was also a small proportion of the total variance attributed to instreamVis and maxrateToQ50, respectively (Figure 26).

Figure 26. Relative importance of predictor variables from linear models for each metric with $R^2 > 0.3$. Metrics are ordered by decreasing R^2 values and stressor predictor variables at listed at the top (i.e. CHLA, instreamVis and maxrateToQ50), followed by environmental predictors. Predictor variables not explaining any variance were excluded during model selection.

Traits

Similarly to the metric models, CHLA had the largest absolute effect size on traits (mean = 0.172, 95% CI = 0.16–0.19), followed closely by instreamVis (0.168, 95% CI = 0.09–0.24) and maxrateToQ50 (0.14, 95% CI = 0.09–0.12). These three stressors were included in five, four and three of the seven best-performing traits models, respectively (Figure 27). As with metric models, the direction and magnitude of these effects was trait-dependent.

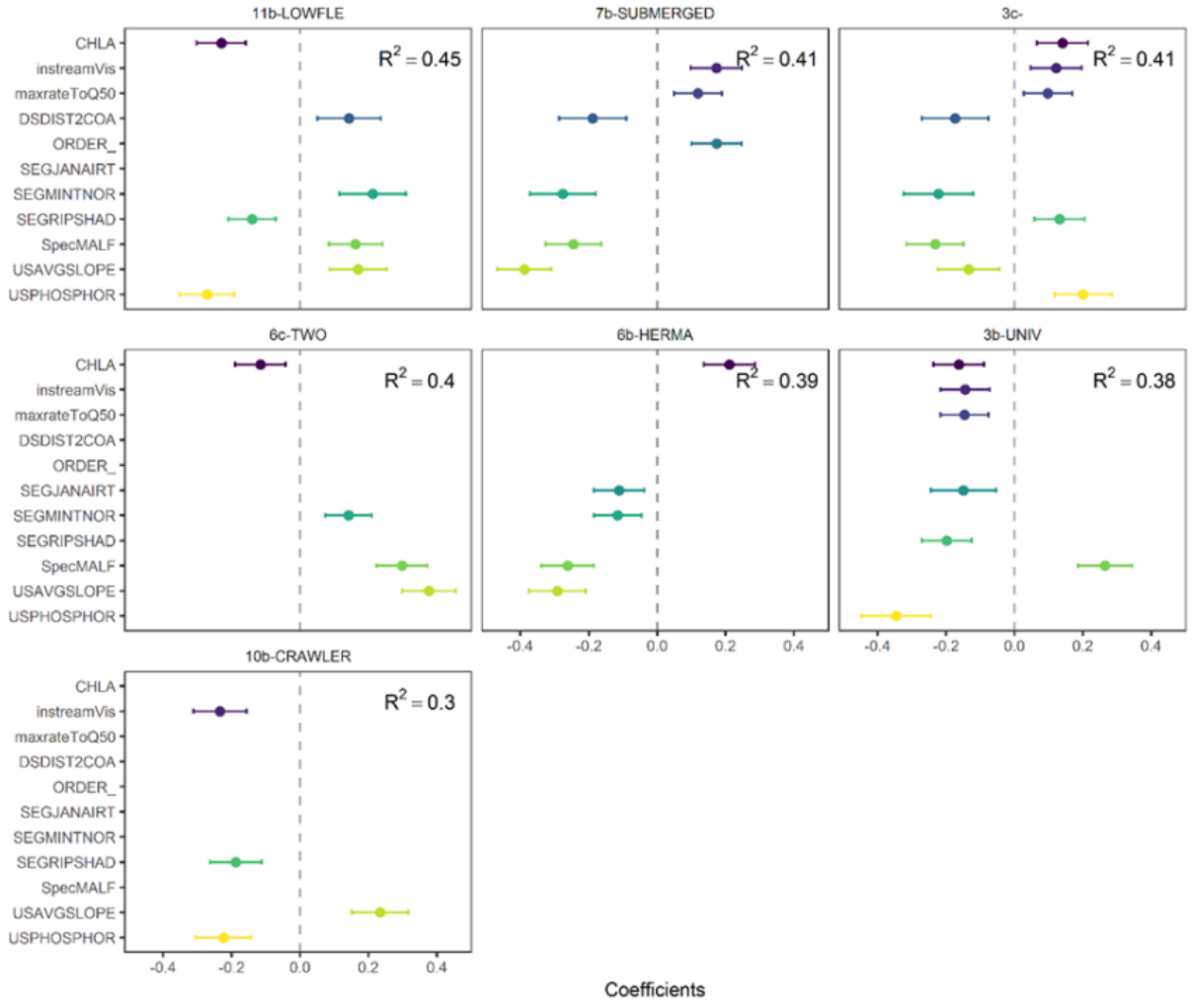


Figure 27. Regression coefficients (± 95% CI) of fixed effects obtained from linear models for each trait with $R^2 > 0.3$. Metrics are ordered by decreasing R^2 values, displayed in each panel. Predictor variables without regression coefficients were excluded during model selection.

Overall, instreamVis was the stressor that had the highest relative importance, explaining on average 20% (± 4% SE) of the total variance (R^2) of trait data. This was followed by CHLA and maxrateToQ50, with average importance of 15% (± 2 SE) and 3% (± 1% SE), respectively (Figure 28). Traits with greatest independent variance

attributable to CHLA alone were 11b-Lowflex, 6c-Two and 6b-Herma (Figure 28). The only trait with variance attributable to instreamVis independently was 10b-Crawler, whereas no trait had independent variance attributable to maxrateToQ50 alone (Figure 28).

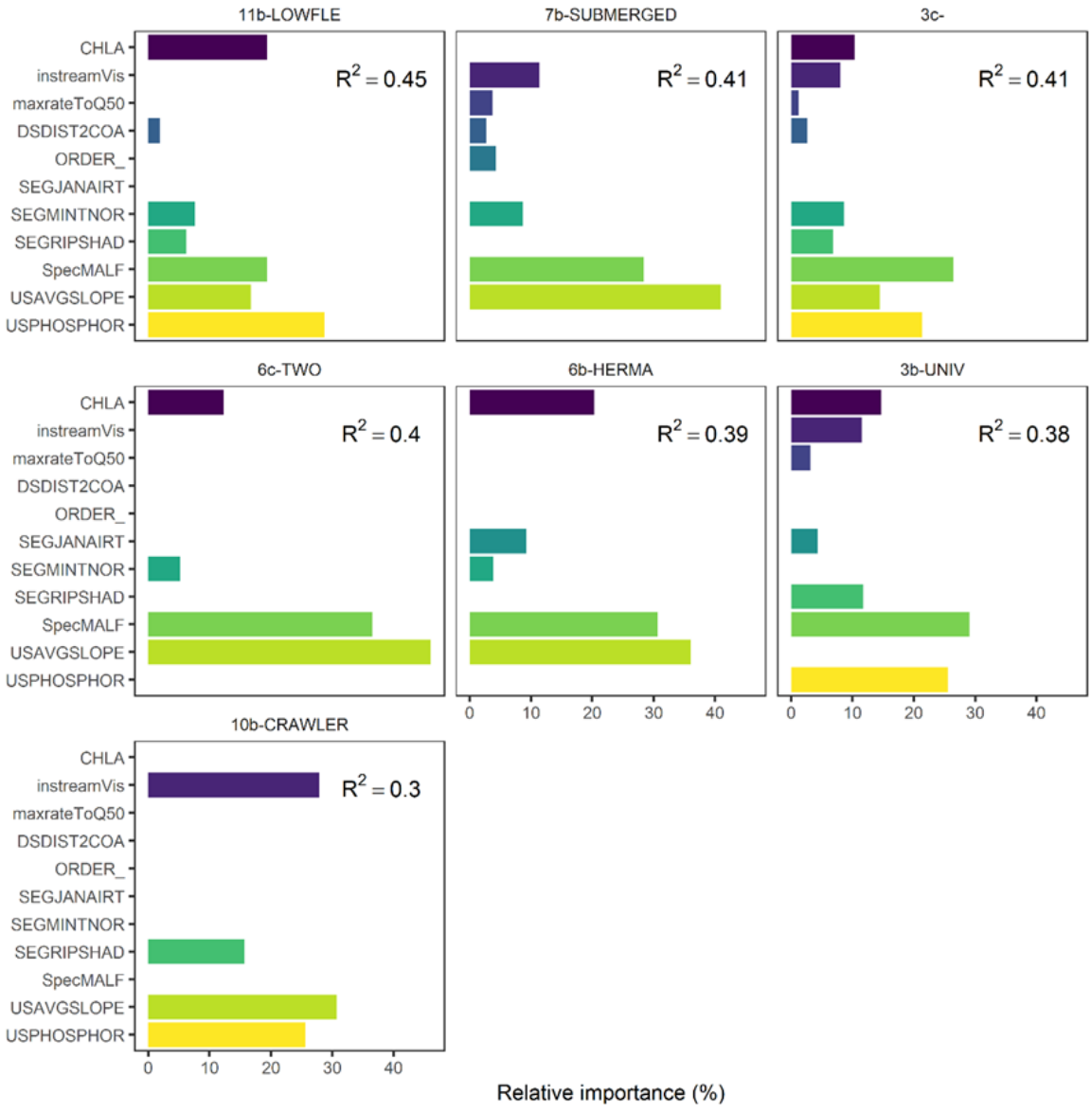


Figure 28. Relative importance of predictor variables from linear models for each trait with $R^2 > 0.3$. Traits are ordered by decreasing R^2 values and stressor predictor variables at listed at the top (i.e. CHLA, instreamVis and maxrateToQ50), followed by environmental predictors. Predictor variables not explaining any variance were excluded during model selection.

5.4.3. Final metric selection

The final selection of seven macroinvertebrate metrics including traits with the strongest links to human land-use pressure gradients based on GF and LM analyses are presented in Table 20. Eight stressor-specific metrics including traits with the strongest links to enrichment and deposited fine sediment based on GF and LM analyses are presented in Table 21.

Table 20. Final selection of macroinvertebrate metrics and traits that had the strongest links with human land-use pressure along with the adjusted R^2, the effect size (coefficient) of the primary stressor in the multiple linear regression model, and response direction (from fitted function of the RF model). *excluding Hydroptilidae; +to both heavy pastoral land use and native vegetation removal (100% native vegetation cover)

Stream health metric	Metric code	R^2	Stressor effect size	Response direction+
MCI (Stark's tolerance values)	MCI_hb	0.50	T1Native = 0.27	negative
EPT richness*	EPTrich*	0.32	T1Native = 0.37	negative
% EPT richness*	pEPTrich*	0.38	T1Native = 0.35	negative
One reproductive cycle per individual	4a-CPI1	0.53	T1Native = 0.20	negative
Female bears eggs in/on body	8c-Protected	0.46	T2PastoralHeavy = 0.14	positive
Oviposition site submerged	7b-Submerged	0.42	T1NativeVeg = -0.22	positive
Adult or larvae stage in water	16b-Aduorlar	0.47	T1NativeVeg = 0.20	negative

Table 21. Final selection of stressor-specific metrics and traits with the strongest links with nutrients or sediment along with the adjusted R^2, the effect size (coefficient) of the primary stressor in the multiple linear regression model, and response direction (from fitted function of the RF model).

Metric	Metric code	R^2	Stressor effect size	Response direction
Richness of enrichment-tolerant taxa	chl_richness_increaser	0.37	Chla = 0.33	positive
'Enrichment MCI'	chl_MCI_like	0.40	Chla = -0.30	negative
Low body flexibility	11b	0.47	Chla = -0.23	negative
Reproduction: hermaphrodism	6b	0.39	Chla = 0.21	positive
Richness of sediment-tolerant taxa	sed_richness_increaser	0.44	Sediment = 0.21	positive
'Sediment MCI'	sed_MCI_like	0.40	Sediment = -0.21	negative
Oviposition site: submerged	7b	0.41	Sediment = 0.17	positive
Attachment to substrate: crawler	10b	0.30	Sediment = -0.23	negative

5.5. Summary

5.5.1. Responses to land-use pressure gradients and the selection of candidate metrics

Macroinvertebrate metrics

The catchment-scale models, with land-use cover as the predictor, were used to select those metrics that best responded to impacts from human land use (90 metrics). After excluding highly correlated metrics, we examined the ecological response shapes and overall effect size of 26 candidate metrics and further reduced the list to 14 candidate metrics, for which we built multiple linear regression models to statistically confirm the link between metrics and land use. Based on output from GF and LM analyses, we selected seven core metrics (Table 20) for consideration of an ecosystem health assessment framework (Section 6).

Linear regression (LM) analysis generally confirmed that native vegetation cover and heavy pastoral land use gradients were much more informative than the exotic vegetation or urban land-use gradients for describing variation in current macroinvertebrate metrics. EPT metrics (richness and % richness) featured among the best metrics across the native vegetation cover gradient while none of the EPT metrics featured among the best metrics across the pastoral land-use gradient.

Traits

We scrutinised trait responses for concordance with ecological theory. Among the nine trait modalities that responded to heavy pastoral land use in the GF analysis, four

responses could be backed by theory. Bearing of the eggs in or on the female's body (8c-Eggsprotected) is favoured with increasing pressure suggesting that protection of the eggs increases the chance for survival. Having only a single reproductive cycle per individual (4a-CPI1), by contrast, reduces the chance of survival compared to individuals that have two or more reproductive cycles. Finally, oviposition of submerged eggs (7b-Submerged) appears to increase chance of survival compared to oviposition of eggs at the surface. However, there was also a case where response direction was opposite to expectations based on theory; a higher number of descendants would be expected to be favoured (as opposed to one descendant, 2a-DESC1) as it increases the chance of survival under increased disturbance from pastoral land uses. As with metrics, gradual responses were also the general response shape for the selected traits, making them suitable metrics for ecosystem health assessment. However, LM variance partitioning showed that on average less variance was attributable to land use than environmental descriptors for traits compared to metrics. Instead, average upstream slope and mean annual low flow were important predictors of macroinvertebrate traits. This suggests that traits may les useful then metrics as general indicators of stream health, unless spatial variation due to environment factors can be fully accounted for.

5.5.2. Responses to stressor gradients

Stressor-specific metrics

The reach-scale models, with stressor gradients as predictors, were used to select those that best responded to proximate stressors (88 metrics). After excluding highly correlated metrics, we examined the ecological response shapes and overall effect size of 28 candidate metrics and further reduced the list to 18 candidate metrics, for which we built multiple linear regression models to statistically confirm the link between metrics and stressors. Based on GF and LM output, we selected eight core metrics (Table 21) for consideration in an ecosystem health framework (Section 6).

Sediment-specific metrics were predominantly among the best 16 metrics across the sediment cover gradient and nutrient-specific metrics across the enrichment gradient in GF analysis. None of the stressor-specific metrics based on abundance featured among the 16 best metrics for each stressor gradient. Also, the Chl MCI and Sediment MCI metrics based on assigning tolerance values according to bins that are equally spread across the raw stressor gradient appeared to perform better than those based on tolerance values assigned according to bins equally spread across the log-transformed stressor gradient. Hence, the latter variants (chl_MCI_like_log and sed_MCI_like_log) were not selected for further analysis. According to LM (R^2 values and effect sizes), the enrichment metrics based on tolerant taxa (increasers) or on tolerant and sensitive taxa (Chla MCI) appeared to be more strongly related to increasing chlorophyll-*a* than those based on sensitive taxa (decreasers). For sediment metrics, there was less of a difference between these tolerant and sensitive taxa metrics.

Stream health metrics

EPT richness and %EPT richness but not %EPT abundance, featured among the best 16 metrics for chlorophyll-*a* and deposited sediment in the GF analysis. The MCI featured among the best 16 for chlorophyll-*a*, but not for sediment. However, effect sizes according to fitted functions of the RF models were generally smaller for these stream health metrics compared to respective stressor-specific metrics. Furthermore, LMs suggested that stream health metrics were related to both stressors while the stressor-specific metrics were mostly related to only their linked stressor and not, or only to a smaller degree, to the other stressor. Overall this confirms that stream health metrics including the MCI, which was developed to respond to enrichment, are less suited to indicate impacts of specific stressors compared to stressor-specific metrics.

Traits

Four of the traits responded across the enrichment gradient according to current ecological theory. Hermaphrodites (6b-Herma) increased while the male-female reproduction trait (6c-Two) decreased, suggesting the hermaphroditic reproductive technique increases the chance of survival under stress. By contrast, a single reproductive cycle per year (3b-Univ) decreases the chance of survival. Finally, low body flexibility (11b-Lowflex) appears to be a disadvantage in enriched streams. However, effect sizes of the fitted functions of the RF models for traits were considerably smaller for the enrichment gradient than those of the stressor-specific metrics. This was also confirmed by LM analysis where environmental descriptors, on average, described more variance than stressors for traits.

Four of the traits responded across the sediment stress gradient according to current ecological theory and these were different from those that responded to enrichment. Crawlers (10b-Crawler) decreased with increasing sedimentation. By contrast, oviposition of submerged eggs (7b-Submerged), the plurivoltine trait (3c-Pluriv) and small body sizes (1a-Size1) increased, suggesting that these traits are favoured under sediment stress. For example, a higher number of reproductive cycles per year should increase the resilience of populations to sedimentation events. Submerged egg layers may relate to those taxa that lay on aquatic plants, hence are resilient to increasing sediment on the benthos.

Overall, while traits appeared to be less strongly linked to enrichment and sediment than the stressor-specific metrics, they may nonetheless be suitable for stream health assessment as they discriminate between the two stressors.

5.5.3. Suitability of stressor-specific metrics as NOF attributes

As expected commonly-used macroinvertebrate indicators of general stream health were responsive to multiple land uses. For example, MCI_hb and %EPT richness (excl. Hydroptilidae) responded predominantly to native vegetation cover but were also responsive to urban and pastoral cover and water abstraction (maxrateToQ50)

(Figure 21). Similarly, these metrics were responsive to multiple proximate stressors based on results of the reach-scale models. The dominant stressor pathway was chlorophyll-*a* for MCI_hb and sediment for %EPT richness, but both also responded to changes in the other stressor, as well as water abstraction (Figure 25). Hence, these metrics are not stressor-specific metrics and indeed more suitable for assessing the general health of streams as they respond to multiple stressors. MCI_hb had the most independent variance attributed to a single stressor (chlorophyll-*a* = 20%) reflecting its original development as an indicator of organic enrichment (Figure 26).

By contrast, several stressor-specific metrics and traits responded to one stressor only and as such change in metric values could be attributed to a change in that primary stressor. For example, the enrichment MCI metric (chl_MCI_like) had a large effect size in response to chlorophyll-*a* and the most independent variance attributed to a stressor, chlorophyll-*a* = 33% (Figure 24). For sediment, 10b-Crawlers had the largest effect size and independent variance attributed to sediment = 28% (Figure 24). These percentages require validation with independent data sets.

Each of the stressor-specific metrics and traits responsive to one stressor could be used to provide evidence of the ecological effects of existing or new ‘attributes’ in the NPS-FM. For example, sediment-specific metrics could be used to help inform a sediment attribute (Depree et al. 2017) and nutrient-specific metrics could be used to validate or update a periphyton attribute.

6. FRAMEWORK FOR ASSESSING ECOSYSTEM HEALTH USING MACROINVERTEBRATES

6.1. Overview

The main aim of this task was to develop a framework for the inclusion of macroinvertebrate metrics in the NPS-FM to assess the Ecosystem Health (EH) value. Recently, in August 2017, amendments to the 2014 NPS-FM (Ministry for the Environment 2017) have taken effect which now include the MCI as a compulsory monitoring tool. However, MCI is a single metric and on its own does not necessarily represent EH. The strength of multiple-metrics for EH assessment and reporting are globally recognised. As such, we aimed to develop a multi-metric and followed an international approach to identify a combination of metrics that represent the key properties of ecosystem health (EH) including organisation/composition, richness/diversity, functional aspects and tolerance.

6.2. Introduction

Defining ecosystem health

Ecosystem health is defined in the NPS-FM as follows: 'In a healthy freshwater ecosystem ecological processes are maintained, there is a range and diversity of indigenous flora and fauna, and there is resilience to change' (MfE 2014). The NPS-FM definition is consistent with the Rapport et al. (1998) definition of ecosystem health as including vigour, organisation, and resilience. According to Rapport et al. (1998), vigour is measured in terms of activity, metabolism or primary productivity (i.e. 'ecological processes' in the NPS-FM. Organisation can be assessed as the diversity and number of interactions between system components. (i.e. 'range and diversity'), and resilience is measured in terms of a system's capacity to maintain structure and function in the presence of stress (i.e. 'resilience to change'). The latter follows the definition of resilience from Holling (1973), as 'a measure of the persistence of systems and of their ability to absorb change and disturbance and still maintain the same relationships between populations or state variables'. In other words, resilience can be considered an emergent property of healthy systems that occurs when a range and diversity of structural and functional components are present.

The NPS-FM definition of EH includes the importance of indigenous communities. The 'nativeness' of faunal community composition has also previously been considered a defining component of ecological integrity (Schallenberg et al. 2011). While New Zealand currently does not have many non-native macroinvertebrate taxa, including some measure of indigenousness (e.g. % non-native taxa) may future-proof any assessment of EH against the effects of subsequent invasions by new species, or increases in abundance or range expansions by existing non-native species due to

climate change etc. Some organisational metrics, such as EPT richness already account for non-native taxa to some degree.

What is missing from the NPS-FM definition of EH is the importance of a reference condition for assessing EH (Bailey et al. 2004). For example, defining the reference condition, such as 'minimally disturbed' provides the benchmark, and the deviation from the benchmark provides a measure of how healthy or unhealthy the ecosystem is.

Overseas examples of frameworks for assessing ecosystem health

'Ecosystem health' is a value managed through freshwater policy in many jurisdictions internationally. Rarely is a single metric used to indicate overall EH. Consistently, benthic macroinvertebrates are used as an element of ecosystem health assessment, and mostly these approaches incorporate multiple macroinvertebrate metrics. The use of multiple metrics recognises that values such as EH have multiple components that allow for a more holistic assessment while retaining the ability to focus in on the effects of specific stressors or on value components. Multi-metric approaches combine several metrics into a single so-called multi-metric index (MMI). There are multiple ways of combining several metrics into an MMI. Metrics can also be assigned to ecosystem components before being combined into an MMI for overall assessment.

In Australia for example, South East Queensland has an EH report card for rivers and streams informed by measures of macroinvertebrates, fish, water quality, nutrient processes and ecosystem processes (Bunn et al. 2010). The macroinvertebrate component is represented by an MMI composed of the three macroinvertebrate metrics SIGNAL[6], Family Richness, and EPT richness. The MMI site score is calculated as follows. First, each metric score is converted to a standardised score by comparing the observed value at a site with a reference condition. Standardised scores range from 0 (maximum observed or deviation from reference condition) to 1 (equal to reference condition). Secondly, standardised metric scores are averaged to create an MMI score also ranging between 0 and 1 (Sheldon et al. 2012).

Focussing specifically on how benthic macroinvertebrates are used to assess EH, in member states of the European Water Framework Directive the biological quality element of ecosystem status is in part assessed using multiple macroinvertebrate metrics that provide measures of **richness/diversity**, **composition/abundance**, **tolerance**, and **functional aspects**. Metrics are used in biological quality assessment at multiple levels. For example, the German system has four levels of metric interpretation (Meier et al. 2006):

1. Ecological Quality Class calculated from a multi-metric index

[6] SIGNAL (Stream Invertebrate Grade Number Average Level) is a simple biotic index calculated from the sum of pollution tolerance scores (1–10) for each taxon divided by the total number of taxa (Chessman BC 2003. New sensitivity grades for Australian river macroinvertebrates. Marine and Freshwater Research 54: 95-103.).

2. causes of degradation such organic pollution, acidification, general degradation
3. results of the core metrics, useful for data interpretation purposes
4. results of all metrics including those not used for the multi-metric index

A similar approach has been adopted in states and territories of the United States to meet the requirements of the Clean Water Act 1972[7]. A suite of macroinvertebrate metrics is used to calculate an MMI, providing measures of richness, composition, tolerance, and trophic status. The few metrics that are consistent across individual states include total taxon richness and EPT richness (Carter & Resh 2013). The macroinvertebrate multi-metric index is only one component of a biological integrity assessment that usually also includes habitat and fish, but may also include periphyton, amphibians, macrophytes and birds.

Existing New Zealand frameworks for assessing ecosystem health

The 2017 amended NPS-FM (MfE 2017) requires councils to establish methods for monitoring the extent at which the EH value is provided for, including at least methods for macroinvertebrate communities and the health of indigenous flora and fauna. The MCI is specifically stated as a compulsory monitoring method. As this project is focussed on macroinvertebrates, here we consider only how macroinvertebrate metrics can be used to contribute to an assessment of ecosystem health.

In New Zealand, a nationally applicable assessment framework for EH does currently not exist, hence the purpose of this project. However, work has been done towards a composite index to describe river condition outlined in Ballantine et al. (2012). It included 5 sub-indices of which one was a macroinvertebrate sub-index. The Average Score Per Metric was recommended as the macroinvertebrate sub-index, calculated from EPT richness, %EPT abundance and MCI (Collier 2008). Metric scores were standardised by dividing by the highest scores observed in the dataset for each. The median of these three standardised scores was calculated to provide the overall ASPM score.

Furthermore, a multi-metric index for predicting the ecological integrity of New Zealand streams was based on predictive modelling of national data sets of water quality, macroinvertebrates, fish and ecosystem process metrics (Clapcott et al. 2014). Metrics were chosen to meet a balance between conceptual inclusiveness by measuring ecological integrity components of pristineness, diversity, nativeness and resilience (Schallenberg et al. 2011) and management focussed indicators (i.e. measures that have been and can be widely adopted and communicated). Spatial regression models which accounted for natural environmental gradients for several macroinvertebrate metrics were explored, but only the MCI metric and a trait

[7]The objective of the 1972 CWA is to '… to restore and maintain the chemical, physical and biological integrity of the Nation's waters.'

describing the number of macroinvertebrates reproducing once in a life-cycle were selected to contribute to the MMI based on good model performance. Standardised O/E (Observed/Expected) scores were calculated for each metric by dividing contemporary scores ('O', in this instance predicted values) by predicted reference scores ('E'). O/E scores were weighted prior to aggregation according to two factors: (1) the strength of the predictive model (average of the ability to explain variation in observed values (model R^2) and predictive accuracy) indicating support of ecological relationships and (2) the size of the training data indicating national representativeness. The two weighted O/E scores for MCI and Cycle (equivalent to CPI1 in our study) were averaged to provide a macroinvertebrate bimetric score.

The 2016 Waikato River Report Card (Williamson et al. 2016) extends the environmental report card concept by including cultural and economic aspects based on bicultural values encapsulated in Te Ture Whaimana—the Vision and Strategy for restoration of the Waikato River. It uses the eight Taura (Maori for 'strands of a rope', i.e., kai (food), water security, ecological integrity, experience, sites of significance, economy, water quality, sites of significance) in an A to D grading system calculated from 64 indicators. Grades are assigned by Taura and overall at two levels of spatial aggregation; (i) report card unit (similar to NPS-FM Freshwater Management Units) and (ii) whole catchment. This system is broader than the EH focus of the current report, but provides an example of methods to aggregate a wide variety of indicators into a simplified message on the health and well-being of a catchment and its people.

The macroinvertebrates indicator (for assessing ecological integrity in the Waikato River) was only applied in river tributaries and included the four metrics of MCI, QMCI, %EPT Density and %EPT Richness. Macroinvertebrate metric site scores were assigned to grades A, B, C or D after Stark and Maxted (2007) and Plafkin et al. (1989) (Table 22). The average metric grade for each reporting unit was calculated by first assigning numeric scores to each of the four alphabetic grades, averaging the numeric score and then assigning the alphabetic grade. Higher weightings were subjectively given to MCI over EPT metrics.

Table 22. Grading macroinvertebrate metrics for the Waikato River Report Card.

Grade	MCI	QMCI	%EPT Density	%EPT Richness
A	> 119	> 6	>70	>70
B	100–119	5–6	51-70	51-70
C	80–99	4–5	25-50	25-50
D	< 80	< 4	<25	<25

6.3. Development of a framework for New Zealand streams

6.3.1. Combining the best of two approaches

There are two different approaches to developing a framework for incorporating macroinvertebrates into an assessment of EH—either a 'general' or 'stressor-specific' approach (Hering et al. 2006). A 'general' framework recognises the components of EH as defined by Rapport et al. (1998), including:

- organisation (i.e. analogous to 'range and diversity' in the NPS-FM definition) e.g. biodiversity, species composition, food web structure
- vigour (i.e. analogous to 'ecological processes' in the NPS-FM definition) e.g. rates of production, nutrient cycling
- resilience (e.g. ability to resist and recover from disturbance).

In contrast, a stressor-specific framework would account for the effects of multiple, specific stressors. Both the U.S. and European countries use a combined approach that incorporates impact-specific metrics (e.g. urban land use, agricultural), if not stressor-specific metrics[8], as indicators of the resilience/tolerance component of EH (Table 23).

Table 23. Example framework for how macroinvertebrate metrics could account for components of ecosystem health. Each component potentially has a different 'score' which could be aggregated by averaging or using the minimum score.

Attribute state	Richness/ diversity	Composition/ organisation	Functional aspects	Resilience/ tolerance	EH multi-metric
Excellent					
Good	Metric 1		Metric 3	Metric 4	Average
Satisfactory		Metric 2			
Poor					

Similarly, we suggest a combined framework would be suitable for the inclusion of one or more macroinvertebrate metrics in the NPS-FM. We envisage a multi-metric index made up of several stressor-specific and/or value-specific metrics. This would provide a single overall score as well as diagnostic metrics beneath it that would help inform management decisions.

[8] There is work being done to identify stressor-specific indices for diagnostic purposes in the United States, but currently this idea is in development and not really rigorously practiced – C.P. Hawkins personal comment.

6.3.2. Overview of our approach to developing a macroinvertebrate multi-metric index for New Zealand wadeable streams

We amended the method of Hering et al. (2006) to develop a macroinvertebrate multi-metric index for New Zealand wadeable streams:

1. *Selection of the most suitable form of a multi-metric.* We decided to develop a framework that includes both generic and specific indicators of the effects of land use on stream health, as opposed to either a 'general' or a 'stressor-specific' approach.
2. *Metric selection*. i. We calculated 110 metrics. ii. Metric selection was based on the relationship of metrics with stressors. We were not able to include metrics based on multivariate prediction (e.g. 'O/E taxa loss') because limited national-scale 'reference' data (i.e. data from minimally disturbed sites) was available preventing the calculation of reliable metric values. Hence, we had 109 candidate metrics iii. We defined two levels of stressor gradients. Firstly, catchment-scale descriptors of anthropogenic impact including % native vegetation cover (inverse to native vegetation removal), % exotic vegetation, % pastoral heavy, % urban and surface water allocation pressure (maxrateToQ50), and secondly, stressors assessed at the reach scale, nutrient enrichment (chlorophyll-a), deposited sediment, DIN and DRP iv. We explored the correlation of metrics and stressors as an initial data check. v. To reduce our set of candidate metrics, we explored their response to our stressor gradients using GF and LM. vi. We selected core metrics based on their relative effect size in response to stressors.
3. *Setting class boundaries.* We rescaled metric values from 0-1 and explored their distribution at reference vs non-reference sites. The value that best discriminated reference from non-reference sites was statistically determined and used to define the upper threshold. The difference between the upper threshold and lowest possible observed value (zero) was used to define thresholds and assign quality classes.
4. *Generation of a multi-metric index*. We assigned the metrics to four different components of EH: functional aspects, diversity/richness, organisation/composition and tolerance. We trialled multiple groupings and selected the combination of core metrics whose combined score best delineated reference from impacted sites in the training dataset.
5. *Interpretation of results.* We explored the resulting multi-metric and component metric scores for all sites in the training dataset.

6.3.3. Details on our approach

Selection of the most suitable form of a multi-metric

We grouped all calculated metrics into a framework which could inform multiple components of ecosystem health (Table 24).

Table 24. Matrix of macroinvertebrate metrics calculated as part of this project assigned to EH components. * there are 59 specific trait modalities as described in Table 4, ** multiple variations available, † not available at the national scale.

Metric	Functional aspects	Diversity/ richness	Organisation/ composition	Tolerance
Productivity	X			
Traits*	X			X
Functional diversity**	X	X		
Feeding diversity**	X	X		
Taxon richness		X		
Diversity**		X		
O/E taxa loss†			X	
EPT**		X	X	X
MCI**				X
Sediment metric**				X
Eutrophication metric**				X
LIFENZ**				X
UCI**				X
AMDI				X

Metric selection

Output from the analyses linking metrics to stressors described in Section 5 was used to select the most suitable 'core' metrics for inclusion in an ecosystem health assessment framework among a large set of candidate metrics. Details on the selection process and statistical methods can be found in Section 5. Briefly, first metrics and traits were chosen based on: their relationship with land use when the direction of response was logical and effect size was greater than competing metrics; the R^2 of the LM was > 0.4; and the land-use stressors explained greater than 10% of the total variation. This included the three metrics EPT richness*, %EPT richness* and MCI_hb, and the three traits 4a (single reproductive cycle per individual), 8c (bearing of the eggs in or on the body) and 16b (adult or larval aquatic stages). Core metrics selected here were representative of the EH components functional aspects, diversity/richness and organisation/composition (Table 25). Secondly, metrics and traits were also chosen that responded specifically to reach-scale stressors in a logical way, with large effect sizes, and R^2 of the LM was > 0.3[9] and the reach-scale stressors explained greater than 10% of the total variation. This included the stressor-specific metrics Sed_MCI, Sed_rich_increasers, Chla_MCI and Chla_rich_increaser as well as the traits 10b (crawlers), 7b (submerged ovipositors), 11b (low body flexibility) and 6b (hermaphrodite). This second set of core metrics represented the tolerance component of EH. Combined these core metrics were taken through to the next stage of MMI development.

[9] Reach-scale LM generally explained less variance in metrics than catchment-scale LM and so we reduced the threshold of importance from 0.4 to 0.3 to consider a selection of tolerance metrics.

Table 25. Core metrics assigned to EH components. *excluding Hydroptilidae.

Metric	Functional aspects	Diversity/ richness	Organisation/ composition	Tolerance
Trait 4a (CPI1)	X			X
Trait 8c (Protected)	X			X
Trait 16b (AduorLar)	X			X
EPT richness*		X		X
%EPT richness*			X	X
MCI_hb				X
Sed_MCI				X
Sed_rich_increasers				X
Trait10b (Crawlers)				X
Trait 7b (Submerged)				X
Chla_MCI				X
Chla_rich_increasers				X
Trait11b (Lowflex)				X
Trait 6b (Herma)				X

Identifying reference distributions

Core metrics were scaled to a range from 0–1 using the formula

$$x = (x - \min(x)) / (\max(x) - \min(x))$$

For metrics (or traits) that increased rather than decreased in response to land-use pressure the scaled value was subtracted from 1 so that for all metrics 0 represented the worst and 1 the best condition. We plotted the distribution of metric values which showed a clear separation between reference (defined as land use of > 85% Native vegetation, < 10% Pastoral heavy and 0% Urban) and non-reference sites for core metrics (Figure 29).

We used logistic regression to identify the MMI threshold that best distinguished reference sites from non-reference sites and provided the highest likelihood that any given site would be assigned correctly as reference or non-reference (Figure 29). Threshold selection was automated through the process of plotting a 'receiver operating characteristic' (ROC) curve (Fawcett 2006). The ROC was generated by calculating for each possible threshold value the rates of true positive classification and false positive classification, and plotting these two rates against each other. As the threshold value is adjusted from one end of its possible range to the other, the two rates change gradually, and they plot as a curved line. The area under this curve (AUC) is greatest when the logistic regression best separates reference from non-reference. Values for the AUC statistic range between 0.5 and 1, with values closer to 1 indicating better discrimination between reference and non-reference. The resulting threshold was used to determine the MMI value that defined the boundary between reference condition and impaired condition (e.g. A/B band threshold in the NPS-FM).

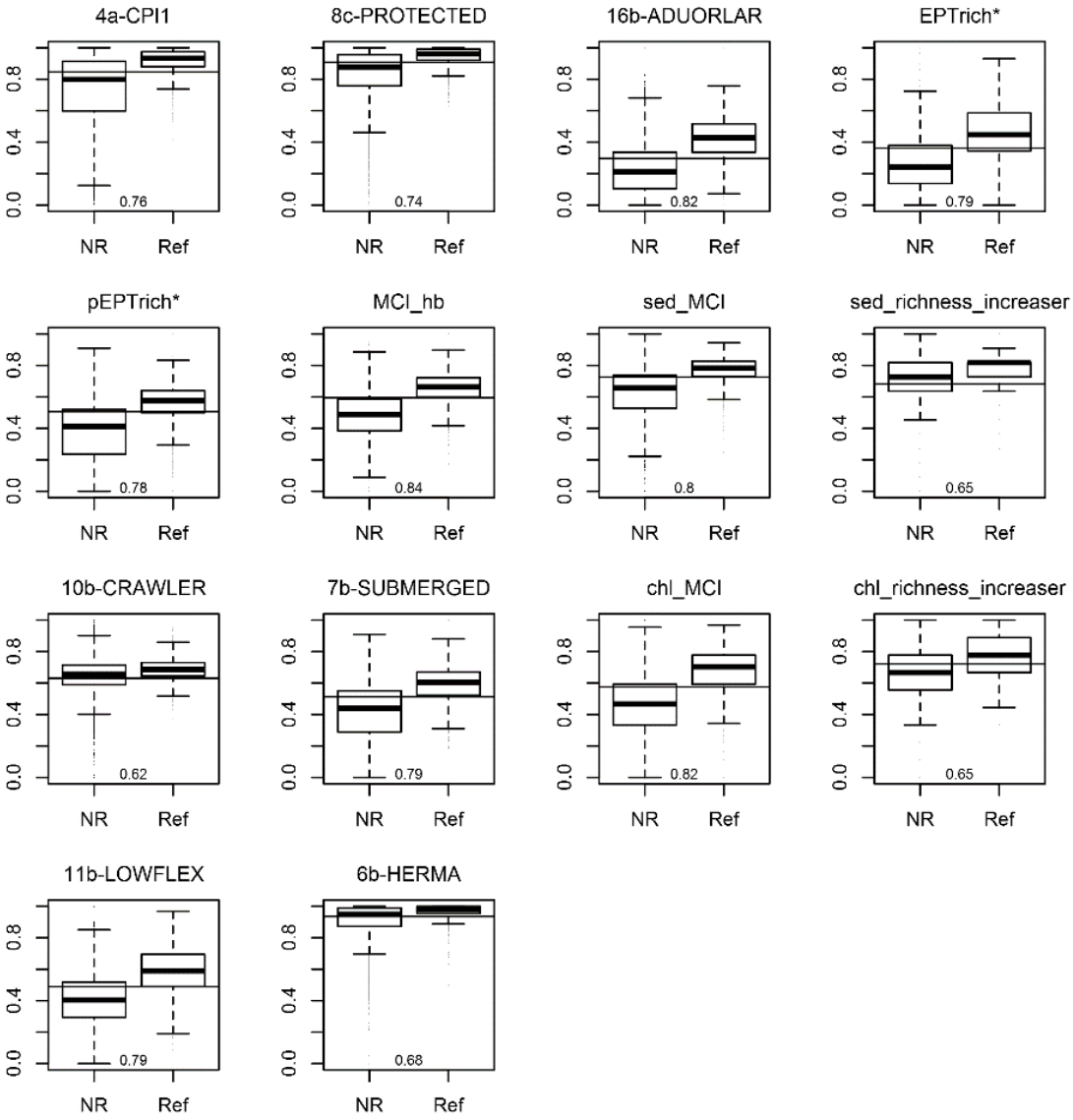


Figure 29. Boxplots of the distribution of metric values at reference (Ref) vs non-reference (NR) sites. The AUC statistic is given and the horizontal line denotes the threshold value that best separates reference from NR.

Generation of a multi-metric index

We trialled numerous combinations of metrics to inform an MMI as recommended by Van Sickle et al. (2010), although we did not take metric correlation into account. Firstly, for each EH component (functional aspect, diversity/richness, organisation/composition, tolerance), we averaged contributing metric values and sequentially removed metrics to determine the optimum number of metrics to discriminate reference from non-reference as measured by the AUC statistic (Table 26). For 'functional aspects', averaging together all three selected core metrics gave an AUC of 0.81, but excluding the trait 8c-Protected improved the AUC to 0.82, although the difference is unlikely to be meaningful. The AUC remained at 0.82 when

16b-Aduorlar was removed. Hence in this case, highest discrimination was achieved with either one or two metrics. For the 'tolerance' component of EH, we treated the stressor-specific sediment and enrichment metrics separately from the general tolerance metrics EPT richness, %EPT richness and MCI_hb.

Table 26. Optimum combination of metrics for each ecosystem health component and for a multi-metric index (MMI) of overall ecosystem health determined by the AUC statistic.

EH component	Contributing metrics	AUC
Functional aspects	All three (4a-CPI1, 8c-Protected,16b-AduorLar)	0.81
	Two (4a-CPI1, 16b-AduorLar)	0.82
	4a-CPI1	0.82
Diversity/richness	EPT richness*	0.79
Organisation/composition	%EPT richness*	0.78
Tolerance - general	All three (MCI_hb, EPT richness*, %EPT richness*)	0.83
	MCI_hb, EPT richness*	0.83
	MCI_hb	0.84
Tolerance - sediment	All four (Sed_MCI, Sed_rich_increasers, 10b-Crawlers, 7b-Submerged)	0.78
	Three (Sed_MCI, 10b-Crawlers, 7b-Submerged)	0.80
Tolerance - chla	All four (Chla_MCI, Chla_rich_increasers, 11b-Lowflex, 6b-Herma)	0.80
	Three (Chla_MCI, 11b-Lowflex, 6b-Herma)	0.83
	Two (Chl_MCI, 11b-Lowflex)	0.83
MMI10	Weighted equally by component	0.83
MMI5	Excluding stressor-specifics and just using MCI_hb for tolerance	0.83
MMI4	Excluding stressor-specifics and just using MCI_hb for tolerance and 4a-CPI1 for functional aspects	0.83

Next, we calculated an MMI by combining the optimum selection of metrics (based on AUC scores) from each EH component by averaging as follows:

MMI10 = Functional aspects ((4a-CPI1 +16b-AduorLar) /2)
+ Diversity/richness (EPT rich*)
+ Organisation/composition (%EPT rich*)
+ Tolerance (MCI_hb + ((Sed MCI +10b-Crawlers + 7b-Submerged)/3)
+ ((Chl_MCI + 11b-Lowflex) /2) /3
/ 4.

The AUC value for the 10-metric combined MMI was greater than or equal to any of the individual metrics or EH components (Table 26, see also Figure 29), except MCI_hb alone. Interestingly, the same high AUC value was observed when we

excluded stressor-specific metrics from the MMI calculation and only used MCI_hb to indicate the tolerance component (i.e. 5-metric MMI), or if we removed 16b-Aduorlar and just used 4a-CPI1 to indicate functional aspects (i.e. 4-metric MMI).

Consequently, a sensitive yet holistic MMI of ecosystem health at the national scale could be achieved by calculating and combining equally all 10 metrics, or just five, or just four (EPT richness*, %EPT richness*, MCI_hb, CPI1,) macroinvertebrate metrics (Figure 30). We did not explore unequal weighting combinations.

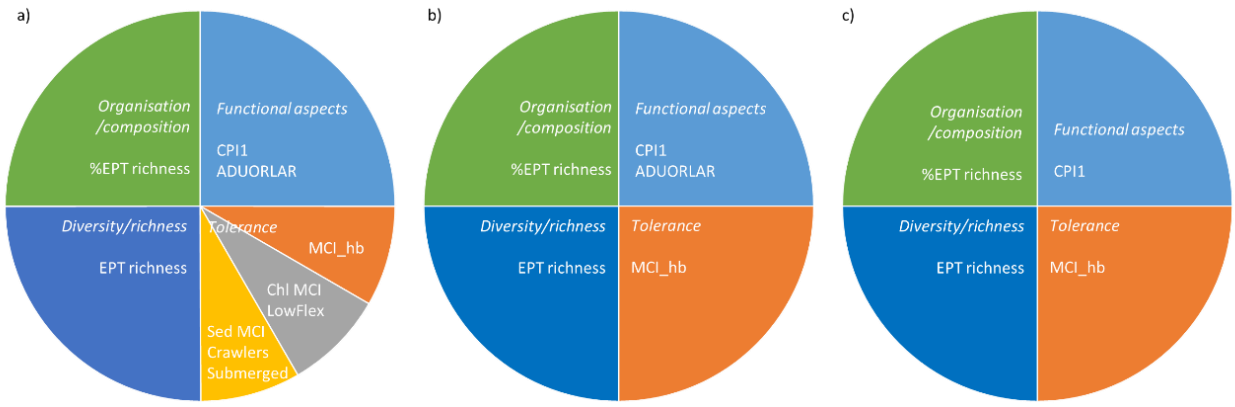


Figure 30. Metrics representing four components (organisation/composition, functional aspects, diversity/richness, tolerance) contributing to a sensitive yet holistic multi-metric index for ecosystem health of wadeable streams using (a) all 10 diagnostic metrics and traits, or (b) a subset that only used MCI_hb to indicate the tolerance component, or (c) a subset that only used MCI_hb to indicate the tolerance component and only one functional aspect metric.

Setting class boundaries

The threshold that best discriminates reference form non-reference sites was used as the upper class boundary for the MMI (0.61). The lowest possible value (i.e. zero) was used as the lowest class boundary (Figure 31). All values in between the upper and lowest boundary were divided into three classes to inform a total of four management classes (i.e. A, B, C, D). Two additional thresholds (0.5, 0.39) divided sample numbers equally so each group had equivalent representation. Thresholds were rounded to the nearest decile, i.e. 0.6, 0.5 and 0.4 (Figure 31). Class A is populated by sites at or close to reference state, whereas classes B, C and D are populated by the equal partitioning of the current SoE dataset. A different dataset (e.g. less or more spatially representative of New Zealand streams) could yield different class thresholds for B/C and C/D. To provide a logical test of the chosen approach to defining classes, we calculated the mean and range of values for the four component metrics (Table 27). While there was overlap in the minimum and maximum component metric values observed in each class, there was a clear reduction in mean values as classes descended from A to D. Looking specifically at the MCI_hb metric, A class streams

had an average MCI score of 129, B class streams had average MCI score of 111, C class streams had average MCI score of 99 and D class streams had an average MCI score of 80 (Table 27).

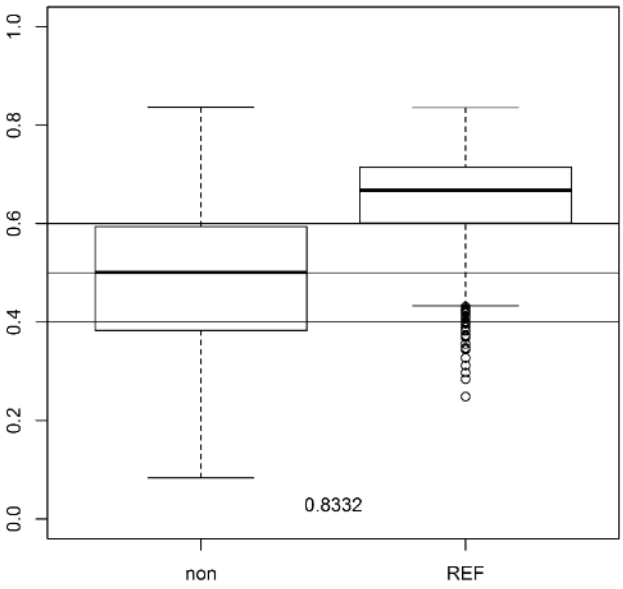


Figure 31. Class boundaries for the MMI based on the discrimination of reference from non-reference sites and a balanced grouping of remaining non-reference sites. Horizontal lines show the thresholds identifying reference from non-reference (0.6) and equal bands (0.5 and 0.4).

Table 27. Mean and range in values for a multi-metric and the four component metrics distributed across four management classes defined by logistic regression (class A) and equal representation (classes B,C and D).

Class	MMI value	CPI1	EPT richness*	%EPT richness*	MCI_hb
A	> 0.6	0.95 (0.7,1)	14 (2, 29)	60 (37, 100)	129 (101, 180)
B	0.5–0.6	0.87 (0.5,1)	9 (1, 19)	47 (15, 75)	111 (86, 165)
C	0.4–0.5	0.76 (0.3,1)	7 (1, 16)	35 (8, 67)	99 (78, 140)
D	< 0.4	0.46 (0,1)	3 (0, 10)	16 (0, 62)	80 (30, 120)

6.3.4. National trial and management guidelines

Temporal variation

We explored the change in MMI scores over time at 243 SoE sites where 10 or more years of data were available. The component metrics with the average lowest coefficient of variation (CV) was 10b-Crawlers followed by MCI_hb, whereas the MMI

had the 3rd lowest CV (Table 28). Both EPT metrics had high inter-annual variation, but 16b-Aduorlar had the highest CV which was more than 3 times the CV of the MMI. The annual CV for MMI4 ranged from 4 to 15 and averaged 7 for 32 sites defined as reference based on land use rules. In contrast, the coefficient of variation ranged from 3 to 53 and averaged 14 for non-reference sites (Table 28). As such, year to year variation was on average twice as high for impacted compared to reference streams.

Table 28. Average percentage coefficient of variation in MMI and component metrics at reference (N = 30) and non-reference (N = 213) State of the Environment sites. * = excluding Hydroptilidae.

Component	Metric	Non-reference sites			Reference sites		
		Mean	Min	Max	Mean	Min	Max
Functional aspect	4a-CPI1	14	2	69	4.5	1.1	13.1
	16b-Aduorlar	40	8	186	18	9	51
Diversity/richness	EPTrich*	38	10	131	23	8	53
Organisation/composition	%EPTrich*	28	7	142	12.7	7	27.4
Tolerance	MCI_hb	9.2	2.9	23.1	6.4	2.6	11.4
	sed_MCI	13	2	67	5.8	2.7	19.8
	10b-Crawlers	8.3	2.7	24.1	5.9	2.8	11.2
	7b-Submerged chl_MCI	15	4	33	17	7	34
		17	5	40	11.3	4.8	28.8
	11b-Lowflex	23	9	69	13.4	8.4	21.2
MMI10		13	4	45	6.9	4	16.2
MMI4		14	3	53	7.0	3.9	15.1

Spatial variation

We calculated the proportion of SoE sites that were assigned to each management class after calculating the 3-yr mean MMI4 score, and recommended narrative guidelines for each class (Table 29). There was a broad spatial distribution in sites assigned to different quality classes nationally (Figure 32).

Table 29. Percentage of sites assigned to each MMI4 management class from the overall training dataset and those from each reference (Ref) and non-reference (NR) sites, respectively.

Class	Description	MMI value	Overall (%)	Ref (%)	NR (%)
A	High quality environment where macroinvertebrate communities are at or close to natural state	> 0.6	24	78	19
B	Low deviation from natural state, likely to be a good quality environment where human activities have caused some loss of sensitive species but sensitive species are common	0.5–0.6	26	18	27
C	High deviation from natural state, likely to be a fair quality environment where moderately-highly tolerant species dominate	0.4–0.5	19	4	20
D	Substantial deviation from natural state, likely to be a poor quality environment where highly tolerant species dominate	< 0.4	31	0	34

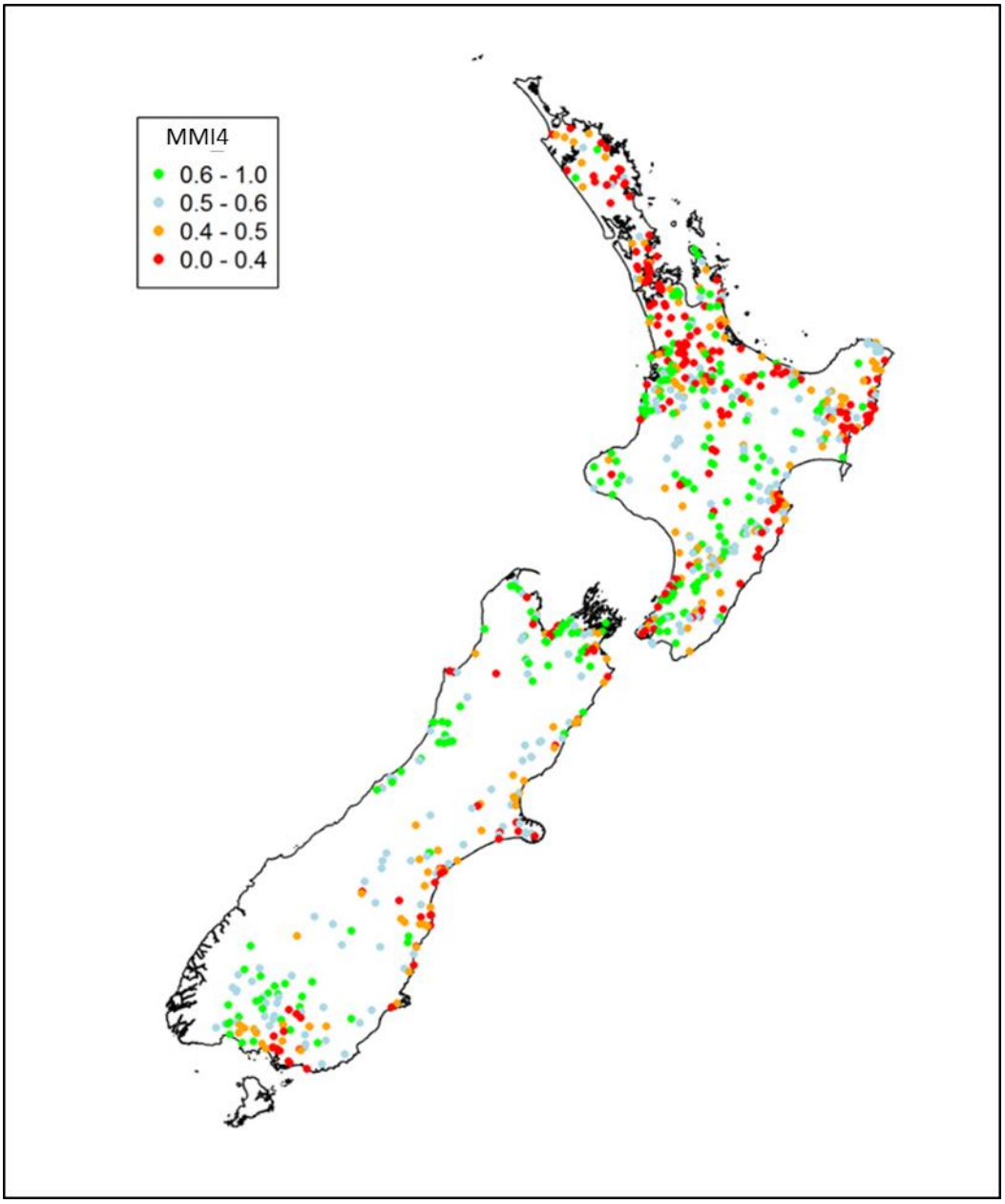


Figure 32. Map of State of the Environment sites assigned to management classes assessed using a multi-metric index based on 4 metrics (MMI4).

We explored the response of the MMI4 to catchment-scale land use predictors and descriptors of environmental variation by developing a random forest model (see Table 15 for description of model predictors). Model leave-one-out cross validation (R^2 = 0.87) was higher than that observed for any component metrics, reported in Section 5. The MMI4 increased in response to native vegetation cover and decreased in response to pastoral heavy cover, which were the 1st and 5th most important predictors, respectively (Figure 33).

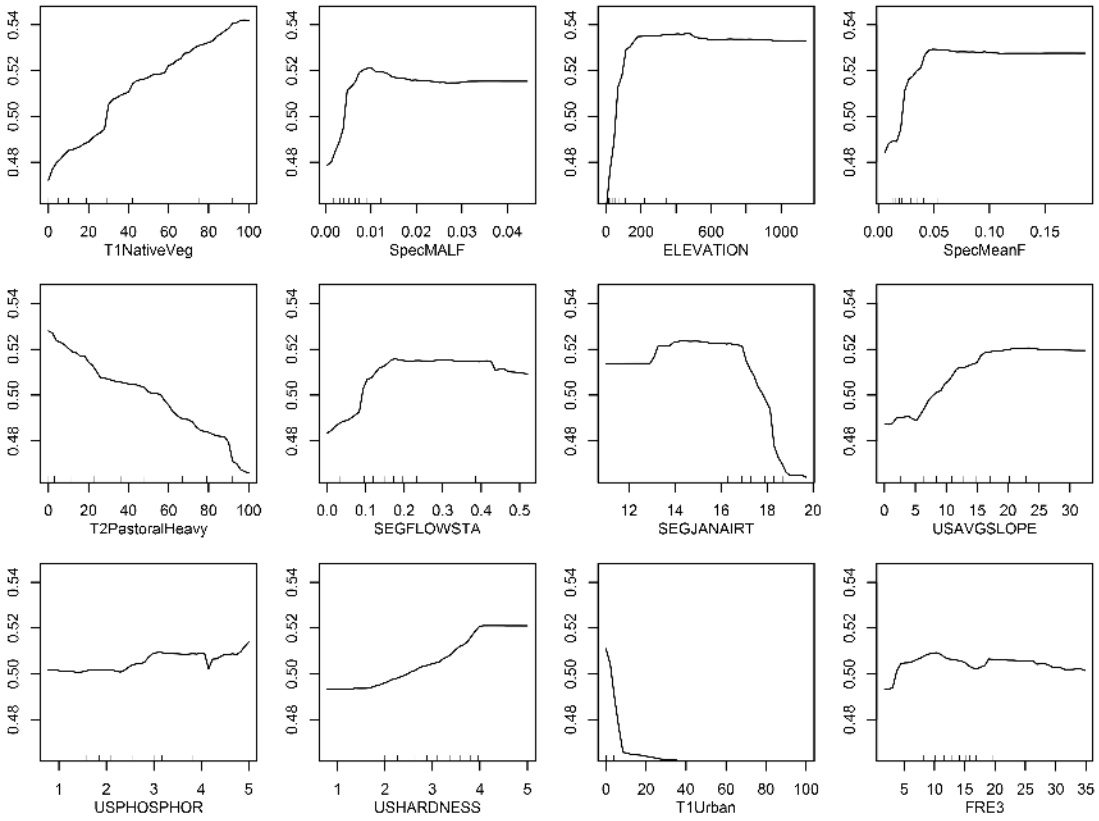


Figure 33. Random forest partial dependence plots of MMI4 in response to the 12 most important land use and environmental predictors.

6.3.5. Discussion

We used an EU-wide approach to developing an MMI. It was similar to that used by member states committed to the European Water Framework Directive in the selection of core metrics to represent EH components (e.g. Vlek et al. 2004). It was also similar to methods previously adopted in North America where component metrics were selected based on their ability to distinguish reference from non-reference sites (e.g. Maxted et al. 2000). The later included five metrics including the number of taxa, number of EPT taxa, % Ephemeroptera, Hilsenhoff Biotic Index, and % clinger mode of existence.

Previously, the development of a multi-metric index has involved scoring component metrics prior to aggregation in the U.S. For example, in a study evaluating macroinvertebrate responses to human activities (Fore et al. 1996), metric values were assigned a score of 5 (similar to expected or reference condition), 3 (different from reference), or 1 (strong deviation from reference). Metric scores were then added to get a final B-IBI score ranging from 11–55. Alternatively, metric values were assigned a continuous score from 0–10 based on linear interpolation between the lowest and highest recorded values and the sum of all contributing metrics divided by

the number of metrics to produce a MMI score between 0–100 (e.g. Hughes et al. 1998). More recently, the selection and scoring of metrics for multi-metric indices has occurred a bit differently in the United States with a greater focus on metric independence (van Sickle 2010; Schoolmaster et al. 2013) and 'deviation from reference' used as a framework to define consistent management classes across naturally variable stream types (Cao et al. 2007; Mazor et al. 2016). In this study, we have used an approach that normalises metric scores, defines classes based on a statistical separation of MMI scores of reference from non-reference sites, then equally divided scores of non-reference sites by frequency of occurrence rather than linear interpolation (which would have resulted in MMI thresholds at > 0.6, 0.4–0.6, 0.2–0.4, < 0.2 and percentage assignment of all SoE sites at 25%, 45%, 22% and 8% for each respective class). Our approach does not currently take the response of the MMI to gradients of natural variation into account.

We developed an MMI based on a combination of metrics representative of four EH components that best statistically distinguished between reference and non-reference, and with approximately 25% overlap in MMI scores between reference and non-reference sites (defined by catchment land use). This overlap reflects the spatial variation in reference condition throughout the country. The MMI could be made more sensitive by defining reference conditions for different stream types (e.g. McDowell et al. 2013; Clapcott et al. 2017). MMI scores could be expressed as observed (e.g. average sites score) divided by expected (e.g. reference condition), which would provide an index for consistent meaning in different settings (Mazor et al. 2016). We trialled an observed/expected MMI score approach (Appendix 6) and it showed similar outputs as our approach reported above. However, the ability to differentiate reference from non-reference sites was lower probably due to weak predictive models for some component metrics. Improved model performance may come from considering confounding factors when developing reference predictions (Schoolmaster et al. 2013).

The temporal variation in some component metrics of our MMI was quite high, especially for EPT metrics, which have a large weighting in the calculation of the MMI and hence probably lead to the MMI having greater temporal variation than the MCI_hb metric. Likewise, Collier (2008) observed greater temporal variation in EPT metrics compared to the ASPM. To retain important component metrics it may be necessary to define a multi-year window for EH assessment, for example a three-year rolling mean for the MMI.

We recommend a two-tiered approach to assess the EH of wadeable rivers in New Zealand. Firstly, a 4-component multi-metric can be used to provide a macroinvertebrate sub-index of stream ecosystem health. The 4-component multi-metric is conceptually holistic, capable of distinguishing reference from non-reference, has moderate temporal variation (varies less at reference sites over time), and has a

stronger predictive relationship with land use than any component metrics. We propose interim management classes for the multi-metric, but recommend further refinement is needed before it can be used in a reference-condition based assessment of EH. Secondly, six functional and tolerance metrics and traits can be used to diagnose why any given site occurs in a management class. Currently, diagnostic metrics and traits have been developed (although they still require validation) for deposited sediment and nutrient enrichment (Section 3). New diagnostic metrics and traits could easily be incorporated into the diagnostic toolbox as they become available.

7. SUMMARY

7.1. Overview of project outputs

The primary objectives of this study were to define the quantitative relationship between macroinvertebrate metrics (new and existing) and human stressors and to explore the connection between macroinvertebrate metrics and the Ecosystem Health (EH) value. In doing so, the applicability of using macroinvertebrate metrics to measure the EH value in the NPS-FM was tested. To address the research objectives the following tasks were undertaken:

- collation of existing data and calculation of existing metrics including updating the macroinvertebrate species traits database (Section 2)
- proof of concept of new stressor-specific metrics (Section 3)
- exploration of a multivariate approach to assessing EH (Section 4)
- characterisation of the quantitative link between metrics and stressors (Section 5)
- development of a framework to include macroinvertebrate metrics in the NPS-FM to assess the Ecosystem Health value (Section 6).

Outputs from the above tasks are summarised here.

7.1.1. Collation of existing data and calculation of metrics

We compiled and taxonomically standardised a national benthic macroinvertebrate database containing two datasets spanning the years 1994 to 2016. The first dataset comprised macroinvertebrate community data from regional councils and the National River Water Quality Network (N = 1,966 sites). The second dataset comprised macroinvertebrate community data and associated physicochemical variables (deposited sediment, suspended sediment, nutrients, periphyton, temperature, dissolved oxygen) from published and unpublished research (N = 973 sites). The compiled database is available from the authors on request.

We calculated 90 existing taxonomic and trait-based metrics describing the benthic macroinvertebrate communities of New Zealand streams. In doing so, we updated the New Zealand macroinvertebrate trait database by incorporating new knowledge from published studies and expert opinion to determine the trait 'profile' of each taxon. The traits database includes affinity scores for sixteen traits with between two and five modalities each, i.e. 59 trait modalities. The traits database is housed by NIWA and trait modalities for currently 495 taxa are available from the authors on request.

7.1.2. Proof of concept of new stressor-specific metrics

We explored different ways to generate tolerance values for the development of stressor-specific metrics:

- a systematic review of the sediment literature using Eco Evidence software
- expert assignment during and following a workshop for sediment, nutrients (via a periphyton pathway), dissolved oxygen, temperature, and metals
- data generation using gradient forest modelling of the research dataset for deposited sediment and nutrient enrichment (via a periphyton pathway).

Tolerance values derived by expert opinion and gradient forest analysis are provided in this report and the Eco Evidence library is available on request.

7.1.3. Exploration of a multivariate approach to assessing EH

Exploration of a multivariate approach to the assessment of macroinvertebrates, whereby a biological classification of reference sites (N = 538) was used to predict the appropriate reference condition for all segments in the training dataset. Based on a new biological classification of sites, we made predictions by constructing a River Invertebrate Prediction and Classification System (RIVPACS) reference condition-type model. The predictive accuracy of the biological classification model was high and similar to that observed for other multivariate models developed overseas. We also tested a multivariate model based on a stream typology—the Freshwater Ecosystems of New Zealand (FENZ) classification of sites, which performed equally well. Output of this work, which represents the early development of a predictive model for assessing macroinvertebrate communities in New Zealand rivers, is provided in this report.

7.1.4. Quantifying the link between metrics and stressors

We explored the relationship between all metrics, taxonomic and trait composition, and measures of catchment condition and proximate stressors using gradient forest and general linear model analyses of the first two datasets combined. The gradient forest analyses were used to identify which metrics had the largest relative effect sizes and consistent response shapes in relation to catchment-scale land use and in-channel proximate stressors. The linear models were used to quantify the relationship between metrics and drivers and identify how much independent variance could be assigned to specific drivers. This latter analysis identified that some of the newly developed stressor-specific metrics could be considered truly stressor-specific, whereas existing tolerance metrics such as the MCI are responsive to multiple stressors and hence good indicators of the multiple impact pathways of land use on stream ecosystem health. A proposed sediment attribute was informed by regression of the sediment-specific metrics developed in this study and a national dataset of measured deposited fine sediment (Depree et al. 2017).

7.1.5. Framework to assess the macroinvertebrate component of stream ecosystem health

We identified a combination of metrics that represent the key properties of EH including organisation/composition, richness/diversity, functional aspects and tolerance, and when combined in a multi-metric index (MMI), best distinguishes reference from non-reference sites. The MMI includes 4 key metrics (CPI1, EPT richness*, %EPT richness*, and MCI_hb) and an additional 6 functional and tolerance metrics provide diagnostic tools for further assessing the pathways through which degradation, or conversely rehabilitation, is occurring. The MMI scores range from 0 to 1 and were grouped into 4 management classes reflecting deviation from reference state:

A – at or similar to natural state
B – low deviation from natural state
C – high deviation from natural state
D – substantial deviation from natural state.

A national trial of the recommended framework using the national dataset showed broad spatial variation in the MMI related to land use with 24% of all SoE sites at or similar to reference state on average in the last 3 years and 31% of all sites at substantial deviation from reference. The inter-annual variation in multi-metric scores was twice as high on average at non-reference sites compared to reference sites (< 85% native vegetation catchments). Of the component metrics, MCI_hb had the lowest inter-annual variation and was the most likely to correctly distinguish reference from non-reference at the national scale.

7.2. Recommendations

7.2.1. Improve stressor-specific metrics and validate those using independent datasets.

We provide proof-of-promise of stressor-specific metrics where macroinvertebrate communities are responsive to instream measurements of periphyton and sediment (Section 3). Metrics were developed using a unique method (GF) to assign tolerance scores to individual taxa. The advantage of the GF approach is that the effect of multiple stressors can be considered simultaneously, identifying the primary stressor for specific taxa. The disadvantage of the GF approach is that it is novel and has not been tested against alternative published approaches to illustrate its effectiveness. It would provide evidence to support the application of currently developed stressor-specific metrics if they were tested, and if possible improved, by exploring alternative ways to assign tolerance values. For example, using the iterative rank procedure used to develop MCI values, but with a measured rather than hypothesised stressor gradient(s).

We did not progress the development of stressor-specific metrics for general habitat degradation, or dissolved oxygen, temperature or heavy metals, because we had insufficient instream measurements of these predictor variables to use the GF approach to assigning tolerance values. However, we did compile expert opinion-based tolerance values for multiple taxa in relation to these stressors. These tolerance values could be used to explore the development of other stressor-specific metrics. The testing of such metrics will be limited until more stressor data become available, and diurnal variability in DO and temperature can be accounted for. Likewise, the suitability of the AMDI to indicate heavy metals remains to be tested nationally.

7.2.2. Improve the multivariate model predicting taxa occurrence

We provided evidence that a national model predicting the probability of occurrence of benthic macroinvertebrate taxa can be developed in New Zealand based on the natural variation observed at reference sites (Section 4). There is scope to improve and validate the proof-of-concept provided here to demonstrate the applicability of a multivariate model for assessing ecosystem health. Improvements could be achieved by:

- a larger reference site training dataset
- optimising model selection including environmental predictors
- evaluation of rare or geographically limited species
- use of abundance data.

The main gains in validating a national multivariate model would be in the provision of (1) a metric that measures the EH component of community composition/organisation, which is currently assessed by the EPT richness metric, and, (2) stream-type-specific benchmarks. Ideally the accuracy and precision of reference benchmarks developed from a multivariate model would be tested against those predicted from models of metrics (see below).

7.2.3. Define site or stream-type specific reference conditions for core metrics to capture spatial variation

There was a significant difference between the distributions of metric values at reference sites compared to non-reference sites at the national scale. However, previous studies have shown that MCI_hb reference conditions can vary spatially (Clapcott et al. 2017). Without a site or stream-type specific reference condition assessments will be biased against sites where high metric values do not naturally occur. We trialed developing random forest models that predict reference benchmarks for varying metrics (Appendix 5). This approach, or a similar boosted regression approach could be refined to provide the best available predictions of reference condition for specific metrics (e.g. Waite et al. 2014; Mazor et al. 2016). During model

refinement, the influence of varying land use rules to define reference state could be explored.

7.2.4. Link between benthic macroinvertebrates and other EH components

We have shown that additional value can be gained from macroinvertebrate monitoring through the identification of metrics that indicate and quantify the effects of specific stressors. Future work includes an analysis of how macroinvertebrates could be used to contribute to an assessment of other components of the stream ecosystem, such as higher order consumers (i.e. fish). For example, studies have shown that size-structured macroinvertebrate data can be used to infer the food suitability for maximum growth of drift feeding fish (Allen 1951; Shearer & Atalah 2015). Exploration of a food-web framework for more holistic assessment of ecosystem health is an option.

8. ACKNOWLEDGEMENTS

We gratefully acknowledge data provided by regional councils via LAWA (www.lawa.org.nz) and the parallel MfE-funded sediment report (Depree et al. 2017). Also, various researchers approved access to their published and unpublished research datasets collected in partnership with project authors; our thanks to Chris Hickey, Phillip Jellyman, Jay Piggott, Francis Magbanua, Elizabeth Graham, Francis Burdon, Robin Holmes, Katie Blackmore, Javad Ramezani, Katharina Lange, Romana Salis, Jan Macher, David Reid, Rebecca Eivers, Cale Riddle, Jason Eden, Aslan Wright-Stow and Colin Townsend.

We thank staff from the Ministry for the Environment for their financial support of this project. In particular, thanks to Alice Bradley for ensuring macroinvertebrates stayed on the MfE radar in terms of the NPS-FM; thanks to Evan Harrison for managing the project and alongside Pattern Reid providing constructive feedback on an early report draft; and thanks to Evan Harrison and Carl Howarth for contributing to a project workshop.

We thank John Stark for providing an updated table of tolerance values for MCI calculation as well as for valuable discussions on metric development. We thank Adam Canning, a PhD student at Massey University, for supporting Russell Death with the calculation of productivity metrics.

We thank Chuck Hawkins of Utah State University for providing a robust and constructive review of this final project report. With his feedback, we are able to prioritise recommendations for the ongoing development of metrics to be included in a national policy for the management of New Zealand's freshwaters.

9. REFERENCES

Allen KR 1951. The Horokiwi stream. A study of the trout population. New Zealand Marine Department Fisheries Bulletin 10.

Armitage P, Moss D, Wright J, Furse M 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research 17: 333-347.

Bailey RC, Norris RH, Reynoldson TB 2004. Bioassessment of freshwater ecosystems using the reference condition approach. Boston, MA, Springer US.

Baker ME, King RS 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods in Ecology & Evolution 1: 25-37.

Ballantine D 2012. Developing a composite index to describe river condition in New Zealand: A composite river condition index for New Zealand. Prepared for Ministry for the Environment. NIWA Client Report No: HAM2012-131.

Baumgärtner D, Rothhaupt K-O 2003. Predictive length–dry Mass regressions for freshwater invertebrates in a pre-alpine lake littoral. International Review of Hydrobiology 88(5): 453-463.

Bêche LA, McElravy EP, Resh VH 2006. Long-term seasonal variation in the biological traits of benthic-macroinvertebrates in two Mediterranean-climate streams in California, USA. Freshwater Biology 51: 56-75.

Bêche LA, Resh VH 2007. Biological traits of benthic macroinvertebrates in California Mediterranean-climate streams: long-term annual variability and trait diversity patterns. Fundamental and Applied Limnology 169(1): 1-23.

Benke AC, Huryn AD, Smock LA, Wallace JB 1999. Length-mass relationships for freshwater macroinvertebrates in North America with particular reference to the southeastern United States. Journal of the North American Benthological Society 18(3): 308-343.

Biggs BJF, Kilroy C 2000. Stream periphyton monitoring manual. Prepared for The New Zealand Ministry for the Environment.

Bonada N, Doledec S, Statzner B 2007. Taxonomic and biological trait differences of stream macroinvertebrate communities between Mediterranean and temperate regions: implications for future climatic scenarios. Global Change Biology 13(8): 1658-1671.

Booker DJ 2010. Predicting width in any river at any discharge. Earth Surface Processes and Landforms 35: 828-841

Booker DJ 2013. Spatial and temporal patterns in the frequency of events exceeding three times the median flow (FRE3) across New Zealand. Journal of Hydrology (New Zealand) 52: 15-39.

Booker DJ, Woods RA 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. Journal of Hydrology 508: 227-239.

Brey T 2010. An empirical model for estimating aquatic invertebrate respiration. Methods in Ecology and Evolution 1(1): 92-101.

Brey T 2012. A multi-parameter artificial neural network model to estimate macrobenthic invertebrate productivity and production. Limnology and Oceanography: Methods 10: 581-589.

Brey T, Müller-Wiegmann C, Zittier ZM, Hagen W 2010. Body composition in aquatic organisms—a global data bank of relationships between mass, elemental composition and energy content. Journal of Sea Research 64(3): 334-340.

Bryce SA, Lomnicky GA, Kaufman PR 2010. Protecting sediment-sensitive aquatic species in mountain streams through the application of biologically based streambed sediment criteria. Journal of the North American Benthological Society 29: 657-672.

Bunn SE, Abal EG, Smith MJ, Choy SC, Fellows CS, Harch BD, Kennard MJ, Sheldon F 2010. Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation. Freshwater Biology 55: 223-240.

Cantilli R, Stevens R, Sweitlik W, Berry W, Kaufman P, Paul J, Spehar R, Cormier S, Norton D 2006. Framework for developing suspended and bedded-sediments (SABS) water quality criteria. Washington, D.C., U.S. Environmental Protection Agency, Office of Water.

Cao Y, Hawkins CP, Olson J, Kosterman MA 2007. Modelling natural environmental gradients improves the accuracy and precision of diatom-based indicators. Journal of the North American Benthological Society 26(3): 566-585.

Carlisle D, Hawkins C 2008. Land use and the structure of western United States stream invertebrate assemblages: predictive models and ecological traits. Journal of the North American Benthological Society 27(4): 986-999.

Carter J, Resh V 2013. Analytical approaches used in stream benthic Macroinvertebrate biomonitoring programs of state agencies in the United States. U.S. Geological Survey Open-File Report, 1129.

Casanoves F, Pla L, Di Rienzo JA, Díaz S 2011. FDiversity: a software package for the integrated analysis of functional diversity. Methods in Ecology and Evolution 2(3): 233-237.

Charvet S, Kosmala A, Statzner B 1998. Biomonitoring through biological traits of benthic macroinvertebrates: perspectives for a general tool in stream management. Archiv für Hydrobiologie 142: 415-432.

Chessman BC 2003. New sensitivity grades for Australian river macroinvertebrates. Marine and Freshwater Research 54: 95-103.

Chevan A, Sutherland M 1991. Hierarchical partitioning. The American Statistician 45(2): 90-96.

Chevene F, Doledec S, Chessel D 1994. A fuzzy coding approach for the analysis of long-term ecological data. Freshwater Biology 31: 295-309.

Clapcott J, Goodwin E 2014. Relationships between MCI and environmental drivers. Prepared for Ministry for the Environment. Ministry for the Environment Report CR 201 and Cawthron Report No. 2507. 21 p.

Clapcott JE, Collier KJ, Death RG, Goodwin EO, Harding JS, Kelly D, Leathwick JR, Young RG 2012. Quantifying relationships between land-use gradients and structural and functional indicators of stream ecological integrity. Freshwater Biology 57(1): 74-90.

Clapcott JE, Goodwin EO, Snelder TH, Collier KJ, Neale MW 2017. Finding reference: a comparison of modelling approaches for predicting benchmarks for macroinvertebrate community metrics. New Zealand Journal of Marine and Freshwater Research 57(1): 44-59.

Clapcott JE, Goodwin EO, Young RG, Kelly DJ 2014. A multi-metric approach for predicting the ecological integrity of New Zealand streams. Knowledge and Management of Aquatic Ecosystems 415, 03.

Clapcott JE, Young RG, Harding JS, Matthaei CD, Quinn JM, Death RG 2011. Sediment assessment methods: protocols and guidelines for assessing the effects of deposited fine sediment on in-stream values. Nelson, New Zealand, Cawthron Institute.

Clarke RT, Furse MT, Wright JF, Moss D 1996. Derivation of a biological quality index for river sites: Comparison of the observed with the expected fauna. Journal of Applied Statistics 23: 311-332.

Clarke RT, Wright JF, Furse MT 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling 160: 219-233.

Coleman BD 1981. On random placement and species-area relations. Mathematical Biosciences 54: 191-215.

Collier K, Clapcott J, Neale M 2014. A macroinvertebrate attribute to assess ecosystem health for New Zealand waterways for the national objectives framework - issues and options. Environmental Research Institute Report 36, University of Waikato, Hamilton.

Collier KJ 2008. Average score per metric: an alternative metric aggregation method for assessing wadeable stream health. New Zealand Journal of Marine and Freshwater Research 42: 367-378.

Collier KJ, Wright-Stow AE, Smith BJ 2004. Trophic basis of production for a mayfly in a North Island, New Zealand, forest stream: contributions of benthic versus hyporheic habitats and implications for restoration. New Zealand Journal of Marine and Freshwater Research 38(2): 301-314.

Coysh J, Norris R 1999. Waikato Region AUSRIVAS / RIVPACS trial. Report prepared for the Ministry for the Environment. SMF Contract No. 5091.

Cuffney TF, Qian SS 2013. A critique of the use of indicator-species scores for identifying thresholds in species responses. Freshwater Science 32(2): 471-488.

Culp JM, Armanini DG, Dunbar MJ, Orlofske JM, Poff NL, Pollard AI, Yates AG, Hose GC 2011. Incorporating traits in aquatic biomonitoring to enhance causal diagnosis and prediction. Integrated Environmental Assessment and Management 7: 187-197.

Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT 2007. Random forests for classification in ecology. Ecology 88: 2783-2792.

Davy-Bowker J, Clarke RT, Johnson RK, Kokes J, Murphy JF, Zahrádková S 2006. A comparison of the European Water Framework Directive physical typology and RIVPACS-type models as alternative methods of establishing reference conditions for benthic macroinvertebrates. Hydrobiologia 566: 91-105.

Death RG, Collier KJ 2010. Measuring stream macroinvertebrate responses to gradients of vegetation cover: when is enough enough? Freshwater Biology 55: 1447-1464.

Depree C, Clapcott J, Booker D, Franklin P, Hickey C, Matheson F, Shelley J, Unwin M, Goodwin E, Mackman J, Rabel H, Wagenhoff A 2017. Development of ecosystem health thresholds for suspended and deposited sediment in New Zealand rivers and streams. Prepared for Ministry for the Environment. 163 p.

Dolédec S, Phillips N, Scarsbrook M, Riley RH, Townsend CR 2006. Comparison of structural and functional approaches to determining landuse effects on grassland stream invertebrate communities. Journal of the North American Benthological Society 25: 44-60.

Doledec S, Statzner B 1994. Theoretical habitat templets, species traits, and species richness: 548 plant and animal species in the upper Rhône River and its floodplain. Freshwater Biology 31(3): 523-538.

Doledec S, Statzner B 2008. Invertebrate traits for the biomonitoring of large European rivers: an assessment of specific types of human impact. Freshwater Biology 53(3): 617-634.

Doledec S, Statzner B, Bournard M 1999. Species traits for future biomonitoring across ecoregions: patterns along a human-impacted river. Freshwater Biology 42: 737-758.

Eden JT 2016. The impact of heavy metals on benthic macroinvertebrate communities in Christchurch's urban waterways. Unpublished MSc thesis. University of Canterbury.

Eivers R 2006. The response of stream ecosystems to riparian buffer width and vegetative composition in exotic plantation forests. Unpublished MSc thesis, University of Canterbury.

Ellis N, Smith SJ, Pitcher CR 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93(1): 156-168.

Extence CA, Balbi DM, Chadd RP 1999. River flow indexing using British benthic macroinvertebrates: A framework for setting hydroecological objectives. Regulated Rivers-Research and Management 15: 543-574.

Extence CA, Chadd RP, England J, Dunbar MJ, Wood PJ, Taylor ED 2013. The assessment of fine sediment accumulation in rivers using macro-invertebrate community response. River Research and Applications 29: 17-55.

Fawcett T 2006. An introduction to ROC analysis. Pattern Recognition Letters 27: 861-874.

Fore LS, Karr JR, Wisseman RW 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. Journal of the North American Benthological Society 15(2): 212-231.

Gerritsen J, Barbour MT, King K 2000. Apples, oranges, and ecoregions: on determining pattern in aquatic assemblages. Journal of the North American Benthological Society 19: 487-496.

Gray DP, Harding JS 2012. Acid Mine Drainage Index (AMDI): a benthic invertebrate biotic index for assessing coal mining impacts in New Zealand streams. New Zealand Journal of Marine and Freshwater Research 46: 335-352.

Greenwood M, Booker D, Stark J, Suren A, Clapcott J 2015. Updating MCI tolerance values for freshwater invertebrate taxa. Prepared for Environment Southland and Hawke’s Bay Regional Council. NIWA Client Report No. CHC2015-008.

Greenwood MJ, Booker DJ, Smith BJ, Winterbourn MJ 2016. A hydrologically sensitive invertebrate community index for New Zealand rivers. Ecological Indicators 61: 1000-1010.

Haase R, Nolte U 2008. The invertebrate species index (ISI) for streams in southeast Queensland, Australia. Ecological Indicators 8: 599-613.

Hawkes H 1998. Origin and development of the biological monitoring working party score system. Water Research 32: 964-968.

Hawkins CP, Norris RH 2000. Performance of different landscape classifications for aquatic bioassessments: introduction to the series. Journal of the North American Benthological Society 19(3): 367-369.

Hawkins CP, Vinson MR 2000. Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessment. Journal of the North American Benthological Society 19: 501-517.

Heino J, Muotka T, Paavola R, Hämäläinen H, Koskenniemi E 2002. Correspondence between regional delineations and spatial patterns in macroinvertebrate assemblages of boreal headwater streams. Journal of the North American Benthological Society 21: 397-413.

Hering D, Feld CK, Moog O, Ofenbock T 2006. Cook book for the development of a Multimetric Index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR projects and related initiatives. Hydrobiologia 566: 311-324.

Hill RA, Fox EW, Leibowitz SG, Olsen AR, Thornbrugh DJ, Weber MH In press. Predictive mapping of the biotic condition of conterminous-USA rivers and streams. Ecological Applications: n/a-n/a

Hilsenhoff WL 1977. Use of arthropods to evaluate water quality of streams. Department of Natural Resources, Madison, Technical Bulletin No. 100.

Holling CS 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4: 1-23.

Holmes R 2007. Cumulative effects of multiple agricultural stressors on invertebrates in pasture streams. Unpublished MSc thesis, University of Otago.

Hopkins CL 1976. Estimate of biological production in some stream invertebrates. New Zealand Journal of Marine and Freshwater Research 10(4): 629-640.

Horrigan N, Baird DJ 2008. Trait patterns of aquatic insects across gradients of flow-related factors: a multivariate analysis of Canadian national data. Canadian Journal of Fisheries and Aquatic Sciences 65: 670-680.

Hughes RM, Kaufmann PR, Herlihy AT, Kincaid TM, Reynolds L, Larsen DP 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences 55: 1618-1631.

Hughes RM, Larsen DP, Omernik JM 1986. Regional reference sites: a method for assessing stream potentials. Environmental Management 10: 629-635.

Huryn AD 1998. Ecosystem-level evidence for top-down and bottom-up control of production in a grassland stream system. Oecologia 115(1-2): 173-183.

Huryn D 1996. Temperature-dependent growth and life cycle of *Deleatidium* (Ephemeroptera: Leptophlebiidae) in two high-country streams in New Zealand. Freshwater Biology 36(2): 351-361.

James MR 1987. Ecology of the freshwater mussel *Hydriella menziesi* (Gray) in a small oligotrophic lake. Archiv für Hydrobiologie 108: 337-348.

Joy MK, Death RG 2003. Biological assessment of rivers in the Manawatu-Wanganui region of New Zealand using a predictive macroinvertebrate model. New Zealand Journal of Marine and Freshwater Research 37: 367-379.

Karr JR 1981. Assessment of biotic integrity using fish communities. Fisheries 6: 21-27.

Kaufman L, Rousseeuw P 1990. Finding groups in data: An introduction to cluster analysis. Wiley, New York.

Khan KS, Kunz R, Kleijnen J, Antes G 2003. Five steps to conducting a systematic review. Journal of the Royal Society of Medicine 96: 118-121.

Lange K, Townsend CR, Matthaei CD 2014. Can biological traits of stream invertebrates help disentangle the effects of multiple stressors in an agricultural catchment? Freshwater Biology 59(12): 2431-2446.

Leathwick J, Snelder T, Chadderton W, Elith J, Julian K, Ferrier S 2011. Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. Freshwater Biology 56: 21-38.

Leathwick J, West D, Gerbeaux P, Kelly D, Robertson H, Brown D, Chadderton W, Ausseil A 2010. Freshwater ecosystems of New Zealand (FENZ) geodatabase. Users guide. Wellington, Department of Conservation.

Leathwick JR, Elith J, Chadderton WL, Rowe D, Hastie T 2008. Dispersal, disturbance and contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species. Journal of Biogeography 35: 1481-1497.

Leathwick JR, Julian K, Smith B 2009. Predicted national-scale distributions of freshwater macro-invertebrates in all New Zealand’s rivers and streams. NIWA Client Report HAM2009-042.

Leathwick JR, West D, Gerbeaux P, Kelly D, Robertson H, Brown D, Chadderton WL, Ausseil A-G 2010. Freshwater Ecosystems of New Zealand (FENZ) Geodatabase www.doc.govt.nz/conservation/land-and-freshwater/freshwater/freshwater-ecosystems-of-new-zealand/

Liaw A, Wiener M 2002. Classification and regression by randomForest. R News 2(3): 18-22.

Liess M, Von der Ohe P 2005. Analyzing effects of pesticides on invertebrate communities in streams. Environmental Toxicology and Chemistry 24: 954-965.

Liess M, Schafer RB, Schriever CA 2008. The footprint of pesticide stress in communities--species traits reveal community effects of toxicants. Science of the Total Environment 406(3): 484-90.

MacNally R 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between–and reconciliation of–'predictive' and 'explanatory' models. Biodiversity and Conservation 9(5): 655-671.

Magurran AG 2004. Measuring Biological Diversity. Oxford, Blackwell Scientific Ltd.

Maxted JR, Barbour MT, Gerritsen J, Poretti V, Primrose N, Silvia A, Penrose D, Renfrow R 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19(1): 128-144.

Mazor RD, Rehn AC, Ode PR, Engeln M, Schiff KC, Stein ED, Gillett DJ, Herbst DB, Hawkins CP 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science 35: 249-271.

McDowell RW, Snelder TH, Cox N, Booker DJ, Wilcock RJ 2013. Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. Marine and Freshwater Research 64(5): 387-400.

McEwan AJ 2015. Kākahi monitoring guide. Prepared for the Wairarapa Moana Wetlands Group. Greater Wellington Regional Council, Wellington.

McGill BJ, Enquist BJ, Weiher E, Westoby M 2006. Rebuilding community ecology from functional traits. Trends in Ecology & Evolution 21: 178-185.

Meier C, Böhmer J, Biss R, Feld C, Haase P, Lorenz A, Rawer-Jost C, Rolauffs P, Schindehütte K, Schöll F, Sundermann A, Zenker A, Hering D 2006. Extension and adaptation of the national assessment system for benthic invertebrates to international requirements. UFOPLAN No. 202 24 223. 198 p.

MfE (Ministry for the Environment) 2014. National Policy Statement for Freshwater Management 2014, Ministry for the Environment. Wellington, New Zealand. Pp. 34.

MfE (Ministry for the Environment) 2015. A guide to the National Policy Statement for Freshwater Management 2014. Ministry for the Environment. Wellington, New Zealand.

MfE (Ministry for the Environment) 2017. Clean Water: 90% of rivers and lakes swimmable by 2040. Ministry for the Environment. Wellington, New Zealand.

Moore SC 1998. A photographic guide to freshwater invertebrates of Taranaki's rivers and streams, Taranaki Regional Council.

Moss D, Wright J, Furse M, Clarke R 1999. A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. Freshwater Biology 41: 167-181.

Neale MW, Rippey B 2008. A comparison of environmental and biological site classifications for the prediction of macroinvertebrate communities of lakes in Northern Ireland. Aquatic Conservation: Marine and Freshwater Ecosystems 18: 729-741.

Nichols S, Webb A, Norris R, Stewardson M 2011. Eco Evidence analysis methods manual: a systematic approach to evaluate causality in environmental science.

Norris RH, Webb JA, Nichols SJ, Stewardson MJ, Harrison ET 2012. Analyzing cause and effect in environmental assessments: using weighted evidence from the literature. Freshwater Science 31: 5-21.

Ode PR, Rehn AC, Mazor RD, Schiff KC, Stein ED, May JT, Brown LR, Herbst DB, Gillett D, Lunde K, Hawkins CP 2016. Evaluating the adequacy of a reference-site pool for ecological assessments in environmentally complex regions. Freshwater Science 35(1): 237-248.

Parkyn SM, Rabeni CF, Collier KJ 1997. Effects of crayfish (*Paranephrops planifrons*: Parastacidae) on in-stream processes and benthic faunas: A density manipulation experiment. New Zealand Journal of Marine and Freshwater Research 31: 685-692.

Parsons M, Thoms MC, Norris RH 2003. Scales of macroinvertebrate distribution in relation to the hierarchical organization of river systems. Journal of the North American Benthological Society 22: 105-122.

Phillips N, Reid D 2012a. Biological trait analysis: application to the Waikato Regional Council Monitoring Programme Document No: 2113749.

Phillips N, Reid D 2012b. Biological traits: Application to the Auckland Council River Ecology Monitoring Programme. Technical Report 2012/001.

Plafkin JL, Barbour MT, Porter KD, Gross SK, Hughes RM 1989. Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. Rapid bioassessment protocols for use in streams and rivers: Benthic macroinvertebrates and fish. United States Environmental Protection Agency.

Poff NL 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. Journal of the North American Benthological Society 16: 391-409.

Poff NL, Olden JD, Vieira NK, Finn DS, Simmons MP, Kondratieff BC 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25: 730-755.

Pollard A, Yuan L 2010. Assessing the consistency of response metrics of the invertebrate benthos: a comparison of trait-and identity-based measures. Freshwater Biology 55(7): 1420-1429.

Quinn JM, Cooper AB 1997. Land-water interactions at Whatawhata, New Zealand: introduction and synthesis. New Zealand Journal of Marine and Freshwater Research 31: 569-577.

Quinn JM, Croker GF, Smith BJ, Bellingham MA 2009. Integrated catchment management effects on runoff, habitat, instream vegetation and macroinvertebrates in Waikato, New Zealand, hill-country streams. New Zealand Journal of Marine and Freshwater Research 43: 775-802.

R Core Team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabeni CF, Doisy KE, Zweig LD 2005. Stream invertebrate community functional responses to deposited sediment. Aquatic Sciences 67: 395-402.

Rapport DJ, Costanza R, McMichael AJ 1998. Assessing ecosystem health. Trends in Ecology and Evolution 13: 397-402.

Reynoldson T, Norris R, Resh V, Day K, Rosenberg D 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16: 833-852.

Schallenberg M, Kelly D, Clapcott J, Death R, MacNeil C, Young R, Sorrell B, Scarsbrook M 2011. Approaches to assessing ecological integrity of New Zealand freshwaters. Science for Conservation 307: 84.

Schoolmaster DR, Grace JB, Schweiger EW, Mitchell BR, Guntenspergen GR 2013. A causal examination of the effects of confounding factors on multimetric indices. Ecological Indicators 29(Supplement C): 411-419

Schuwirth N, Kattwinkel M, Stamm C 2015. How stressor specific are trait-based ecological indices for ecosystem management? Science of the Total Environment 505: 565-572.

Shearer KA, Atalah J 2015. Influence of the Taharua River on the upper Mohaka River food base for adult trout. Prepared for Hawke's Bay Regional Council. Cawthron Report No. 2731. 27 p.

Sheldon F, Peterson EE, Boone EL, Sippel S, Bunn SE, Harch BD 2012. Identifying the spatial scale of land use that most strongly influences overall river ecosystem health score. Ecological Applications 22(8): 2188-2203.

Smith AJ, Bode RW, Kleppel GS 2007. A nutrient biotic index (NBI) for use with benthic macroinvertebrate communities. Ecological Indicators 7: 371-386.

Snelder T, Biggs B, Weatherhead M 2004. New Zealand River Environment Classification User Guide. Ministry for the Environment by the National Institute of Water and Atmospheric Research. 145 p.

Snelder TH, Biggs BJF 2002. Multiscale river environment classification for water resources management. Journal of the American Water Resources Association 38: 1225-1239.

Southwood T 1977. Habitat, the templet for ecological strategies? Journal of Animal Ecology 46: 337-365.

Stark JD 1985. A macroinvertebrate community index of water quality for stony streams Water and Soil Miscellaneous Publication 87: 1-52.

Stark JD 1998. SQMCI: A biotic index for freshwater macroinvertebrate coded-abundance data. New Zealand Journal of Marine and Freshwater Research 32(1): 55-66.

Stark JD, Boothroyd IKG, Harding JS, Maxted JR, Scarsbrook MR 2001. Protocols for sampling macroinvertebrates in wadeable streams. New Zealand Working Group Report No.1. Prepared for the Ministry for the Environment. Sustainable Management Fund Project No. 5103. 57 p.

Stark JD, Maxted JR 2007a. A user guide for the Macroinvertebrate Community Index. Prepared for the Ministry for the Environment. Cawthron Report No.1166. 56 p.

Stark JD, Maxted JR 2007b. A biotic index for New Zealand's soft-bottomed streams. New Zealand Journal of Marine and Freshwater Research 41: 43-61.

Statzner B, Beche LA 2010. Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems? Freshwater Biology 55: 80-119.

Statzner B, Hildrew AG, Resh VH 2001. Species traits and environmental constraints: entomological research and the history of ecological theory. Annual Review of Entomology 46: 291-316.

Stoffels RJ, Karbe S, Paterson RA 2003. Length-mass models for some common New Zealand littoral-benthic macroinvertebrates, with a note on within-taxon variability in parameter values among published models. New Zealand Journal of Marine and Freshwater Research 37(2): 449-460.

Suren A, Snelder T, Scarsbrook M 1998. Urban Stream Habitat Assessment method (USHA). NIWA Client Report No. CHC98/60.

Thienemann A 1918. Lebensgemeinschaft und Lebensraum. Naturwissenschaftliche Wochenschrift 17: 282-290.

Towers DJ, Henderson IM, Veltman CJ 1994. Predicting dry weight of New Zealand aquatic macroinvertebrates from linear dimensions. New Zealand Journal of Marine and Freshwater Research 28(2): 159-166.

Townsend CR, Hildrew AG 1994. Species traits in relation to a habitat templet for river systems. Freshwater Biology 31: 265-275.

Townsend CR, Uhlmann SS, Matthaei CD 2008. Individual and combined responses of stream ecosystems to multiple stressors. Journal of Applied Ecology 45: 1810-1819.

van Sickle J 2010. Correlated metrics yield multimetric indices with inferior performance. Transactions of the American Fisheries Society 139(6): 1802-1817.

van Sickle J, Hawkins CP, Larsen DP, Herlihy AT 2005. A null model for the expected macroinvertebrate assemblage in streams. Journal of the North American Benthological Society 24: 178-191.

Vlek HE, Verdonschot PFM, Nijboer RC 2004. Towards a multimetric index for the assessment of Dutch streams using benthic invertebrates. Hydrobiologia 516: 173-189.

Wagenhoff A, Clapcott JE, Lau KEM, Lewis GD, Young RG 2017b. Identifying congruence in stream assemblage thresholds in response to nutrient and sediment gradients for limit setting. Ecological Applications 27: 469–484.

Wagenhoff A, Liess A, Pastor A, Clapcott JE, Goodwin EO, Young RG 2017a. Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams. Freshwater Science 36: 178-194.

Wagenhoff A, Shearer K, Clapcott J 2016. A review of benthic macroinvertebrate metrics for assessing stream ecosystem health. Prepared for Environment Southland. Cawthron Report No. 2852. 49 p. plus appendices.

Wagenhoff A, Townsend CR, Phillips N, Matthaei CD 2011. Subsidy-stress and multiple-stressor effects along gradients of deposited fine sediment and dissolved nutrients in a regional set of streams and rivers. Freshwater Biology 56(9): 1916-1936.

Waite IR, Kennen JG, May JT, Brown LR, Cuffney TF, Jones KA, Orlando JL 2014. Stream macroinvertebrate response models for bioassessment metrics: addressing the issue of spatial scale. PLoS ONE 9(3): e90944.

Webb AJ, Miller KA, King EL, de Little SC, Stewardson MJ, Zimmerman JKH, LeRoy Poff N 2013. Squeezing the most out of existing literature: a systematic re-analysis of published evidence on ecological responses to altered flows. Freshwater Biology 58: 2439-2451.

Webb JA, Miller KA, Stewardson MJ, de Little SC, Nichols SJ, Wealands SR 2015. An online database and desktop assessment software to simplify systematic reviews in environmental science. Environmental Modelling & Software 64: 72-79.

Williamson B, Quinn J, Williams E, van Schravendijk-Goodman C 2016. 2016 Pilot Waikato River Report Card: Methods and Technical Summary. Prepared for Waikato River Authority. NIWA Client Report No: HAM2016-011.

Winterbourn MJ 1996. Life history, production and food of *Aphrophila neozelandica* (Diptera: TipuIidae) in a New Zealand stream. Aquatic Insects 18(1): 45-53.

Winterbourn MJ, Gregson KL, Dolphin CH 1989. Guide to the aquatic insects of New Zealand, Entomological Society of New Zealand Auckland.

Wright-Stow AE, Winterbourn MJ 2003. How well do New Zealand's stream-monitoring indicators, the Macroinvertebrate Community Index and its quantitative variant, correspond? New Zealand Journal of Marine and Freshwater Research 37: 461-470.

Yuan LL 2007. Using biological assemblage composition to infer the values of covarying environmental factors. Freshwater Biology 52(6): 1159-1175. Doi: 10.1111/j.1365-2427.2007.01744.x

Zweig LD, Rabeni CN 2001. Biomonitoring for deposited sediment using benthic invertebrates: a test on four Missouri streams. Journal of the North American Benthological Society 20: 643-657.

Zuur AF, Ieno EN, Elphick CS 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1(1): 3-14.

10. APPENDICES

Appendix 1. Tolerance values used for metric calculations in Section 2.

Table A1.1. Tolerance values used for MCI metric calculations, for both hard-bottomed (hb) and soft-bottomed streams. The table is an updated version of the table provided in (Stark & Maxted 2007a). Minor updates have been provided by John Stark (personal communication, January 2017).

Taxon	HB	SB
COELENTERATA		
Hydra	3	1.6
PLATYHELMINTHES	3	0.9
RHABDOCOELA	3	0.9
BRYOZOA	-	4.0
NEMATODA	3	3.1
NEMATOMORPHA	3	4.3
NEMERTEA	3	1.8
OLIGOCHAETA	1	3.8
POLYCHAETA	-	6.7
HIRUDINEA	3	1.2
TARDIGRADA	-	4.5
CRUSTACEA		
Amphipoda	5	5.5
Cladocera	5	0.7
Copepoda	5	2.4
Halicarcinus	-	5.1
Helice	-	6.6
Isopoda	5	4.5
Mysidae	-	6.4
Ostracoda	3	1.9
Paracalliope	5	-
Paraleptamphopus	5	-
Paranephrops	5	8.4
Paranthura	-	4.9
Paratya	5	3.6
Tanaidacea	4	6.8
INSECTA		
Ephemeroptera		
Acanthophlebia	7	9.6
Ameletopsis	10	10.0
Arachnocolus	8	8.1
Atalophlebioides	9	4.4
Austroclima	9	6.5
Austronella	7	4.7
Coloburiscus	9	8.1
Deleatidium	8	5.6
Ichthybotus	8	9.2
Isothraulus	8	7.1
Mauiulus	5	4.1
Neozephlebia	7	7.6
Nesameletus	9	8.6
Oniscigaster	10	5.1
Rallidens	9	3.9

Taxon	HB	SB
Odonata		
Aeshna	5	1.4
Anisoptera	5	6.0
Antipodochlora	6	6.3
Austrolestes	6	0.7
Hemianax	-	1.1
Hemicordulia	5	0.4
Ischnura	-	3.1
Procordulia	6	3.8
Uropetala	5	0.4
Xanthocnemis	5	1.2
Hemiptera		
Anisops	5	2.2
Diaprepocoris	5	4.7
Microvelia	5	4.6
Saldidae	5	3.9
Sigara	5	2.4
Coleoptera		
Antiporus	5	3.5
Berosus	5	-
Copelatus	5	3.7
Dytiscidae	5	0.4
Elmidae	6	7.2
Enochrus	5	2.6
Hydraenidae	8	6.7
Hydrophilidae	5	8.0
Liodessus	5	4.9
Onychohydrus	5	-
Podaena	8	-
Ptilodactylidae	8	7.1
Rhantus	5	1.0
Scirtidae	8	6.4
Staphylinidae	5	6.2
Neuroptera		
Kempynus	5	-
Diptera		
Aphrophila	5	5.6
Austrosimulium	3	3.9
Calopsectra	4	-
Ceratopogonidae	3	6.2
Chironomidae	2	3.8
Chironomus	1	3.4
Corynoneura	2	1.7
Cryptochironomus	3	-

Taxon	HB	SB
Neolimnia	3	5.1
Nothodixa	4	9.3
Orthocladiinae	2	3.2
Paradixa	4	8.5
Paralimnophila	6	7.4
Parochlus	8	-
Paucispinigera	6	7.7
Pelecorhynchidae	9	-
Peritheates	7	-
Podonominae	8	6.4
Polypedilum	3	8.0
Psychodidae	1	6.1
Scatella	7	-
Sciomyzidae	3	3.0
Stictocladius	8	-
Stratiomyidae	5	4.2
Syrphidae	1	1.6
Tabanidae	3	6.8
Tanypodinae	5	6.5
Tanytarsini	3	4.5
Tanytarsus	3	-
Thaumaleidae	9	8.8
Tipulidae	5	3.4
Zelandotipula	6	3.6
Trichoptera		
Alloecentrella	9	-
Beraeoptera	8	7.0
Confluens	5	7.2
Conuxia	8	-
Costachorema	7	7.2
Cryptobiosella	9	-
Diplectrona	9	-
Ecnomidae	8	-
Ecnomina	8	9.6
Edpercivalia	9	6.3
Helicopsyche	10	8.6
Hudsonema	6	6.5
Hydrobiosella	9	7.6
Hydrobiosis	5	6.7
Hydrochorema	9	-
Hydropsyche - Aoteapsyche	4	6.0
Hydropsyche - Orthopsyche	9	7.5
Kokiria	9	-
Neurochorema	6	6.0

Table A1.1, continued

Taxon	HB	SB
Siphlaenigma	9	-
Tepakia	8	7.6
Zephlebia	7	8.8
Plecoptera		
Acroperla	5	5.1
Austroperla	9	8.4
Cristaperla	8	-
Halticoperla	8	-
Megaleptoperla	9	7.3
Nesoperla	5	5.7
Spaniocerca	8	8.8
Spaniocercoides	8	-
Stenoperla	10	9.1
Taraperla	7	8.3
Zelandobius	5	7.4
Zelandoperla	10	8.9
Megaloptera		
Archichauliodes	7	7.3

Taxon	HB	SB
Lepidoptera		
Hygraula	4	1.3
Collembola	6	5.3
ACARINA	5	5.2
ARACHNIDA		
Dolomedes	5	6.2
MOLLUSCA		
Ampullariidae	3	1.6
Glyptophysa = Physastra	5	0.3
Gundlachia = Ferrissia	3	2.4
Gyraulus	3	1.7
Hyridella = Echyridella	3	6.7
Latia	3	6.1
Lymnaeidae	3	1.2
Melanopsis = Zemelanopsis	3	1.9
Physa = Physella	3	0.1
Potamopyrgus	4	2.1
Sphaeriidae	3	2.9

Taxon	HB	SB
Culex	3	-
Culicidae	3	1.2
Diptera indet.	3	2.9
Diptera		
Dixidae	4	7.1
Dolichopodidae	3	8.6
Empididae	3	5.4
Ephydridae	4	1.4
Eriopterini	9	7.5
Harrisius	6	4.7
Hexatomini	5	6.7
Limnophora	3	4.5
Limonia	6	6.3
Lobodiamesa	5	7.7
Maoridiamesa	3	4.9
Mischoderus	4	5.9
Molophilus	5	6.3
Muscidae	3	1.6
Nannochorista	7	-
Neocurupira	7	-

Taxon	HB	SB
Oecetis	6	6.8
Oeconesidae	9	6.4
Olinga	9	7.9
Trichoptera		
Oxyethira	2	1.2
Paroxyethira	2	3.7
Philorheithrus	8	5.3
Plectrocnemia	8	6.6
Polyplectropus	8	8.1
Psilochorema	8	7.8
Pycnocentrella	9	-
Pycnocentria	7	6.8
Pycnocentrodes	5	3.8
Rakiura	10	-
Synchorema	9	-
Tiphobiosis	6	9.3
Triplectides	5	5.7
Triplectidina	5	-
Zelandoptila	8	7.0
Zelolessica	10	6.5

Table A1.2. Revised tolerance values used for MCI metric calculations provided by Greenwood et al. (2015).

Taxon	MCI-hb2	Taxon	MCI-hb2	Taxon	MCI-hb2	Taxon	MCI-hb2
Cnidaria	2	*Austronella*	3	Diptera		*Pycnocentrella*	9
Hydra	2	*Coloburiscus*	8	*Austrosimulium*	6	*Pycnocentria*	6
Platyhelminthes	4	*Deleatidium*	8	Blephariceridae	9	*Pycnocentrodes*	6
Temnocephala	7	*Ichthybotus*	8	*Neocurupira*	9	*Rakiura*	7
Cura	5	*Isothraulus*	8	*Peritheates*	9	*Triplectides*	3
Neppia	8	*Mauiulus*	5	Ceratopogonidae	6	*Triplectidina*	8
Nematoda	4	*Neozephlebia*	6	Ceratopogoninae	5	*Zelandoptila*	4
Nematomorpha	6	*Nesameletus*	7	Chironomidae	5	*Zelolessica*	7
Nemertea	2	*Oniscigaster*	6	Chironominae	4	Ecnomidae	4
Oligochaeta	4	*Rallidens*	6	Chironomini	4	*Helicopsyche*	8
Lumbricidae	5	*Siphlaenigma*	7	*Chironomus*	1	*Hudsonema*	4
Polychaeta	5	*Tepakia*	2	*Harrisius*	4	*Hydrobiosella*	8
Scolecolepides	4	*Zephlebia*	5	*Paucispinigera*	2	Hydrobiosidae	8
Hirudinea	2	Plecoptera		*Polypedilum*	2	*Costachorema*	8
Alboglossiphonia	1	*Acroperla*	6	Tanytarsini	4	*Edpercivalia*	6
Amphipoda	3	*Austroperla*	8	*Tanytarsus*	6	*Hydrobiosis*	7
Chiltonia	2	*Cristaperla*	9	Diamesinae	6	*Hydrochorema*	8
Gammaridae	2	*Megaleptoperla*	7	*Lobodiamesa*	5	*Neurochorema*	7
Orchestia	4	*Nesoperla*	5	*Maoridiamesa*	6	*Psilochorema*	7
Paracalliope	3	*Spaniocerca*	7	Orthocladiinae	4	*Tiphobiosis*	8
Paracorophium	1	*Spaniocercoides*	8	*Corynoneura*	2	Hydropsychidae	8
Paraleptamphopidae	4	*Stenoperla*	8	*Cricotopus*	5	*Aoteapsyche*	7
Paraleptamphopus	4	*Taraperla*	7	*Naonella*	1	*Orthopsyche*	7
Phreatogammarus	3	*Zelandobius*	7	*Pirara*	7	Hydroptilidae	3
Talitridae	3	*Zelandoperla*	7	*Stictocladius*	8	*Oxyethira*	3
Isopoda	3	Megaloptera		Podonominae	6	*Paroxyethira*	2
Austridotea	3	*Archichauliodes*	8	*Parochlus*	8	Oeconesidae	4
Phreatoicidae	3	Odonata		Tanypodinae	4	*Oeconesus*	4
Phreatoicus	2	*Aeshna*	1	Culicidae	1	*Zelandopsyche*	9
Cladocera	1	Anisoptera	1	*Culex*	1	Polycentropodidae	5
Daphnia	1	*Antipodochlora*	1	Dixidae	3	*Plectrocnemia*	7
Simocephalus	1	*Austrolestes*	1	*Nothodixa*	5	*Polyplectropus*	3
Copepoda	1	*Hemicordulia*	2	*Paradixa*	2	Lepidoptera	
Cyclopoida	1	*Ischnura*	1	Dolichopodidae	5	*Hygraula*	1
Ostracoda	3	*Procordulia*	1	Empididae	4	Mecoptera	
Herpetocypris	4	*Xanthocnemis*	1	Ephydridae	4	*Nannochorista*	5
Tanaidacea	4	Hemiptera		Brachydeutera	5	Acarina	4
Amarinus	1	*Anisops*	1	*Ephydrella*	1	*Arrenurus*	2
Helice	1	*Diaprepocoris*	1	*Scatella*	2	*Hydrachna*	1
Hemigrapsus	2	*Hydrometra*	1	Muscidae	4	Limnesiidae	4
Paranephrops	1	*Mesovelia*	1	*Limnophora*	3	Oribatida	3

Table A1.2, continued

Taxon	MCI-hb2
Paratya	3
Mysidae	1
Tenagomysis	1
Mollusca	
Ferrissia	4
Glyptophysa	1
Gyraulus	2
Hyridella	4
Latia	6
Melanopsis	3
Nucula	4
Physella	2
Potamopyrgus	4
Lymnaeidae	1
Austropeplea	3
Pseudosuccinea	2
Sphaeriidae	2
Sphaerium	2
INSECTA	
Ephemeroptera	
Acanthophlebia	8
Ameletopsis	8
Arachnocolus	4
Atalophlebioides	4
Austroclima	5

Taxon	MCI-hb2
Mesoveliidae	1
Microvelia	1
Saldidae	2
Saldula	3
Sigara	2
Coleoptera	
Dytiscidae	1
Antiporus	1
Huxelhydrus	2
Lancetes	2
Liodessus	1
Rhantus	1
Elmidae	7
Gyrinidae	5
Hydraenidae	7
Homalaena	8
Orchymontia	8
Hydrophilidae	3
Berosus	4
Enochrus	1
Ptilodactylidae	7
Scirtidae	6
Neuroptera	
Kempynus	6
Sisyra	2

Taxon	MCI-hb2
Pelecorhynchidae	6
Psychodidae	3
Sciomyzidae	2
Neolimnia	2
Stratiomyidae	2
Tabanidae	7
Tanyderidae	6
Thaumaleidae	5
Tipulidae	8
Aphrophila	8
Eriopterini	8
Hexatomini	5
Limoniinae	8
Limonia	4
Molophilus	6
Paralimnophila	3
Zelandotipula	2
Trichoptera	
Alloecentrella	8
Beraeoptera	9
Confluens	8
Diplectrona	9
Oecetis	3
Olinga	8
Philorheithrus	6

Taxon	MCI-hb2
Piona	1
Zelandobates	4

Table A1.3. Tolerance values used for UCI metric calculations published by Suren et al. (1998).

Taxon	UCI tolerance value
COELENTERATA	
Hydra	-0.607
PLATYHELMINTHES	-0.290
NEMATODA	-0.533
NEMERTEA	1.166
OLIGOCHAETA	-0.277
HIRUDINEA	-0.416
CRUSTACEA	
Amphipoda	-0.567
Cladocera	-0.181
Copepoda	0.077
Isopoda	0.578
Mysidae	-1.051
Ostracoda	-0.670
Paracalliope	0.650
Paranephrops	0.774
Paratya	1.453
INSECTA	
Ephemeroptera	
Austroclima	2.067
Coloburiscus	1.871
Deleatidium	1.161
Mauiulus	2.151
Nesameletus	1.920
Zephlebia	1.890
Plecoptera	
Acroperla	1.184
Austroperla	2.052
Megaleptoperla	1.554
Zelandobius	1.728
Zelandoperla	0.923
Megaloptera	
Archichauliodes	1.729
Odonata	
Austrolestes	-0.766
Hemicordulia	0.797
Xanthocnemis	0.350
Hemiptera	
Anisops	-0.039
Microvelia	-0.169
Sigara	-0.985

Taxon	UCI tolerance value
Diptera	
Austrosimulium	1.026
Ceratopogonidae	-0.611
Chironomini	0.137
Culex	-0.210
Diamesinae	-0.005
Empididae	0.425
Ephydridae	-0.066
Brachydeutera	*-0.541*
Eriopterini	1.028
Hexatomini	0.355
Limonia	-0.026
Muscidae	0.025
Nothodixa	0.911
Orthocladiinae	0.438
Paradixa	-0.365
Paralimnophila	0.248
Psychodidae	0.105
Sciomyzidae	0.383
Stratiomyidae	0.116
Tanyderidae	0.964
Tanypodinae	-0.797
Tanytarsini	0.871
Zelandotipula	0.311
Trichoptera	
Costachorema	0.444
Hudsonema	0.704
Hydrobiosis	0.989
Hydropsyche - Aoteapsyche	1.358
Neurochorema	1.415
Oecetis	-0.772
Oeconesus	0.119
Olinga	2.073
Oxyethira	0.248
Polyplectropus	0.145
Psilochorema	0.571
Pycnocentria	1.462
Pycnocentrodes	1.472
Tiphobiosis	-0.749
Triplectides	0.720
Lepidoptera	

Table A1.3, continued

Taxon	UCI tolerance value
Coleoptera	
Antiporus	-1.079
Elmidae (adults)	2.063
Homeodytes	-0.430
Hydraenidae	1.744
Hydrophilidae	0.704
Liodessus	-0.601
Ptilodactylidae	1.254
Rhantus	0.315
Scirtidae	0.624
Staphylinidae	0.505

Taxon	UCI tolerance value
Collembola	-0.150
ACARINA	0.132
MOLLUSCA	
Gundlachia = Ferrissia	0.343
Gyraulus	-0.565
Latia	1.233
Lymnaea	0.721
Physa = Physella	-0.494
Potamopyrgus	0.023
Sphaeriidae	-0.612
Pisidium	0.688

Appendix 2. Overview of Eco Evidence and a systematic review of sediment effects literature reported in Section 3.

The Eco Evidence framework adopted in this study consisted of eight steps (Norris et al. 2012) that were used to assess evidence on the effect of sediment on macroinvertebrates in the causal criteria analysis:

1. *Problem definition*. Many anthropological activities degrade terrestrial and riparian environments in such a way that they increase the amount of fine sediment found in streams and rivers. Freshwater macroinvertebrates are sensitive to levels of both fine sediment suspended in the water column and deposited on the benthos, with the direct and indirect addition of anthropogenic sediment affecting habitat and food availability, as well as their direct biological functioning.
2. *Research question*. 'What are the effects of anthropogenic sedimentation on macroinvertebrates in freshwater systems?'
3. Conceptual model. Figure A2.1.
4. *Cause-effect hypotheses.* Entries consisted of a term (an entity) and an attribute (a property of the entity), which were structured 'term (attribute)' e.g., *Deleatidium* (abundance). Classifications (drop down lists) were then used to assign hypothesised trajectories of both the cause and effect terms. From the conceptual model, the identified causes were an increase in deposited and suspended sediment and the measures used to quantify them (e.g., percentage cover of fine sediment), whilst the identified effects were a change in both hypothesised sensitive and non-sensitive individual taxa, as well as changes in more general community structure indicators.
5. *Review literature and extract evidence*. A search for all combinations of cause and effect terms was primarily conducted on Web of Science and Google Scholar. Reference lists of relevant studies and those of previous narrative reviews, along with lists of studies that had cited papers with evidence items relevant to any of the hypotheses were also reviewed. Studies were only included if they generated primary data (to eliminate the risk of double counting a data set), and to avoid misinterpretation by citing authors. Furthermore, only studies that proved statistical significance (or insignificance) of evidence items were retained (as guided by Norris et al. 2012).
6. *Revise*. Both the cause-effect hypotheses and conceptual model were revised throughout the analysis as more causes and effects were discovered in the literature, with these being added to the analysis.
7. *Catalogue and weight the evidence*. A total of 65 studies with varying numbers of evidence items were found that were relevant to the ecological effect of fine sediment addition on macroinvertebrates, and were entered into the software for analysis. The weight of evidence assigned to the item was determined from the experimental design and the level of sample replication. These components were summed to give an overall study weight (Table A2.1) with greater weighting

assigned to research having study design that controlled confounding influences and had greater replication of both controls and treatments.

8. *Assess the level of support for the research question*. In the weighting of evidence items, three causal criteria were used to test for a potential cause-effect relationship. These were: **Response** (the presence of a response), **Dose Response** (if a response is present whether there is a dose relationship between the cause and effect), and **Consistency of Association** (the same results amongst numerous studies) (Nichols et al. 2011). High levels of evidence for the Response and Dose Response criteria display an association between the cause and effect, with this occurring when the summed weight for an evidence item is ≥ 20. A summed weight < 20 shows a low level of evidence for the Response and Dose Response criteria. This means as few as three studies with a high quality, robust design may provide enough evidence to support a cause-effect hypothesis, whereas seven poorly-designed studies may not (Norris et al. 2012). This association was only developed into support for a causal link if high Consistency of Association for the cause-effect hypothesis existed as well. For this the weighting of all the studies that *did not* support the hypothesised cause-effect linkage were summed, and if the summed value was ≥ 20, this was considered to indicate lack of consistency and hence low support for causality. A value < 20 therefore indicated high consistency of association and a high level of support for causality (Nichols et al. 2011). The three causal criteria were then collated for each cause-effect relationship to see the level of support for the hypotheses under investigation.

After an evidence item had been weighted, its trajectory was then compared to that of the cause-effect linkage to assess if it contributed to supporting or refuting the hypothesis. When this had been done for all linkages in relevant citations, the weighting values for all evidence items that supported the hypotheses were summed, as were those refuting it. These two totals were then compared to a threshold value (again with a default of 20 summed points) to see the overall strength and direction of evidence, thus reaching one of four conclusions for the hypothesis (Table A2.1) (Webb et al. 2013).

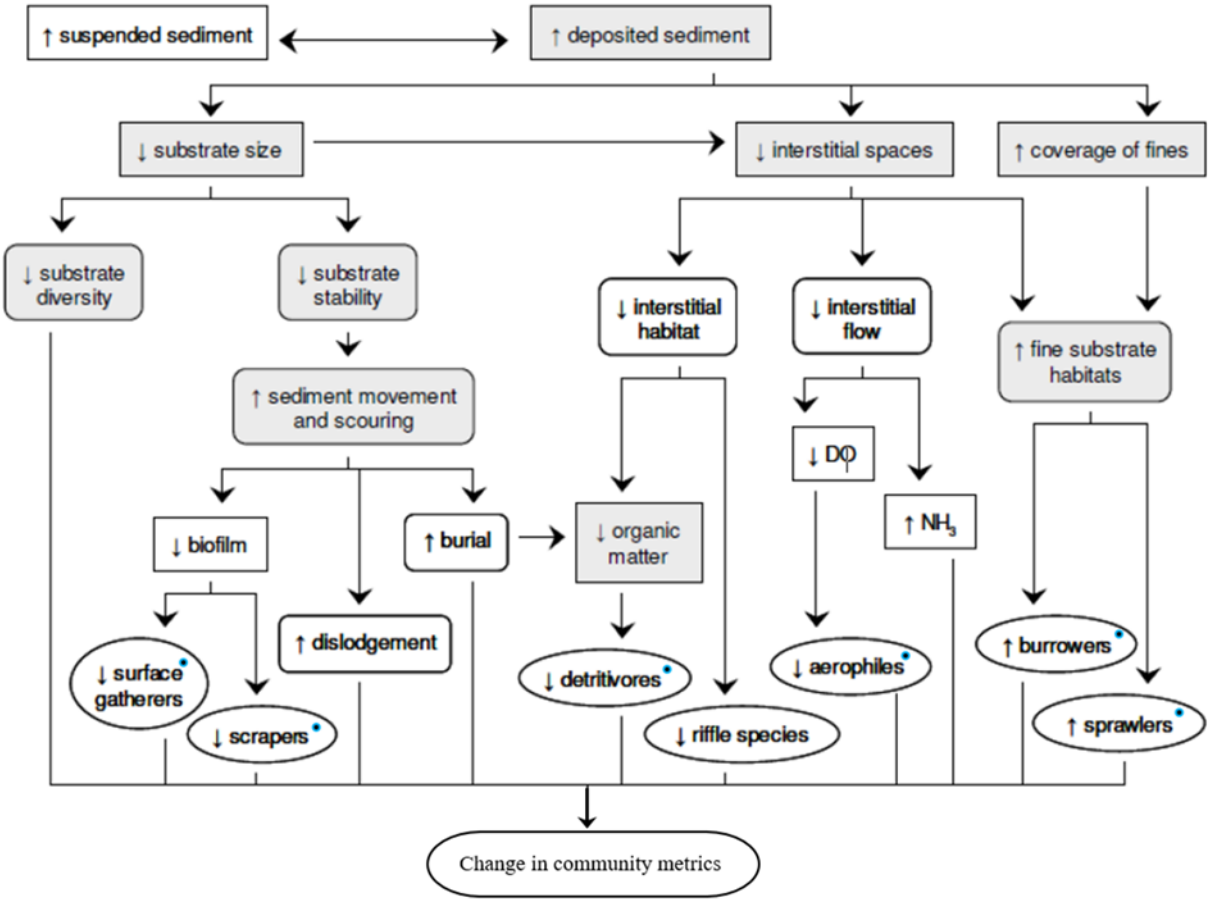


Figure A2.1. Conceptual model for the effect of fine sediment on macroinvertebrates.Rectangular boxes are used for stressors, rounded rectangular boxes show an additional step in the causal pathway, and ovals are used for responses. Responses with blue-black dots indicate individual species are included within these responses. Image adapted from Cantilli et al. (2006).

Table A2.1. The weightings of the different components of an evidence item. Each evidence item consists of a study design weighting and then a weighting for the number of controls and treatments used, except for gradient response studies that instead weigh using the replication of gradient-response models. From Nichols et al. (2011).

Study design component	Weight
Study design type	
After impact only	1
Reference/control vs impact with no before data	2
Before vs after with no reference/control location(s)	2
Gradient response model	3
BACI/ BARI, MBACI, or beyond MBACI*	4
Replication of factorial designs	
Number of reference/control sampling units	
0	0
1	2
> 1	3
Number of impact/treatment sampling units	
1	0
2	2
> 2	3
Replication of gradient-response models	
< 4	0
4	2
5	4
> 5	6

* M = multiple, B = before, A = after, C = control, R = reference, I = impact

Table A2.2. The four possible outcomes of the Eco Evidence Causal Criteria Analysis.

Conclusion	Weighting Supporting Hypothesis	Weighting Refuting Hypothesis	Implications
Support for Hypothesis	≥ 20	< 20	The evidence is consistent with the hypothesis
Support for Alternate Hypothesis	< 20	≥ 20	The evidence falsifies the hypothesis
Inconsistent Evidence	≥ 20	≥ 20	The evidence falsifies the hypothesis, though a subset of the hypothesis may be supported
Insufficient Evidence	< 20	<2 0	There are too few data to test the hypothesis and may also indicate a literature gap

Appendix 3. Supplementary material related to stressor-specific metric development reported in Section 3.

Table A3.1. The mode of sensitivity scores informed by expert knowledge during and following the workshop on taxa sensitivity (Section 3.4). A = highly sensitive, B = moderately sensitive, C = moderately insensitive/tolerant, D = highly insensitive/tolerant/favoured.

Taxon	Deposited sediment	Nutrients	Oxygen	Temp	Metals
COELENTERATA					
Hydra	D	A	B/D	B/D	
PLATYHELMINTHES	C	D	D	B	A
RHABDOCOELA	D	D	D	D	
BRYOZOA	D	A/D	C/D	C/D	
NEMATODA	D	D	D	D	
NEMATOMORPHA	A	A	A	A/C	A
NEMERTEA	C/D	B/D	C/D	C/D	
OLIGOCHAETA	D	D	D	D	C
Tubificidae/Naididae	D	D	D	D	C
POLYCHAETA	C	A/C	B/C	B/C	
HIRUDINEA	C	C	C/D	D	B
TARDIGRADA					
CRUSTACEA					
Amphipoda	C	C	C	C	B
Cladocera	C	C/D	C	C	A
Copepoda	C	C/D	C	C/D	A
Halicarcinus	C	A/C	A/C	B/C	
Helice					
Isopoda	C	A/C/D	B	B	
Mysidae	D	C/D	C/D	D	
Ostracoda	C/D	D	C	C	
Paracalliope	C	C	B	C	B
Paraleptamphopus	C	A/C	A/C	A	
Paranephrops	C	B	B/C	B/C	A
Paranthura					
Paratya	C	C	C	C	B
Tanaidacea					
INSECTA					
Ephemeroptera					
Acanthophlebia	B	A	A	A	
Ameletopsis	A	A	A	A	A
Arachnocolus	B/C	A	A	A	
Atalophlebioides	A	A	A	A	
Austroclima	B	B	B	B	A
Austronella	B	B	A/B	A/B	
Coloburiscus	A	A	A	A	A
Deleatidium	A	B	A	A	A
Ichthybotus	D	A/B/C	B	A	A
Isothraulus	B/C	A	B	A/B	
Mauiulus	A/B	A/B	B	B	
Neozephlebia	B	B	B	B	A
Nesameletus	A	B	A	A	A
Oniscigaster	A/B	B	A	A	A
Rallidens	A	B	A	A	A
Siphlaenigma	A/C/D	A	A	A	A
Tepakia	C	A	A/B	B	
Zephlebia	B	B	B	B	A
Plecoptera					
Acroperla	A/B	A	A	A	D
Austroperla	A	A	A	A	C

Table A3.1, continued

Taxon	Deposited sediment	Nutrients	Oxygen	Temp	Metals
Cristaperla	A	A	A	A	
Halticoperla	A	A	A	A	
Megaleptoperla	A	A	A	A	C
Nesoperla	A	A	A	A	C
Spaniocerca	A	A	A	A	C
Spaniocercoides	A	A	A	A	D
Stenoperla	A/B	A	A	A	C
Taraperla	A	A	A	A	D
Zelandobius	B	B	B	A/B	C
Zelandoperla	A	A	A	A	A/C
Megaloptera					
Archichauliodes	B	B/C	C	C	A/C
Odonata					
Aeshna	B	A/C	B/C	B/C	C
Anisoptera	B/C	A/C	B/C	B/C	C
Antipodochlora	B	A/C	B/C	A/B/C	C
Austrolestes	B	A/C	C	B/C	C
Hemianax	B	A/C	B/C	B/C	C
Hemicordulia	B	A/C	B/C	B/C	C
Ischnura	B	A/C/D	B/C/D	B/C/D	C
Procordulia	B/C	A/C	B/C	B/C	C
Uropetala	A/B	A/C	B/C	A/C	C
Xanthocnemis	C	C	C/D	D	C
Hemiptera					
Anisops	D	D	D	C/D	C
Diaprepocoris	B/D	C/D	C/D	C/D	C
Microvelia	B/D	D	D	D	C
Saldidae	B/D	D	D	D	C
Sigara	D	D	D	C/D	C
Coleoptera					
Antiporus	B/D	B/C	B/C	C	B
Berosus	B/D	B/C	B/C	C	B
Copelatus	B/C	B/C	B/C	B/C/D	B
Dytiscidae	B/C/D	B/C	B/C	C	B
Elmidae	B/C	B/C	A/C	B/D	A/B
Enochrus	B	B/C	B/C	C	B
Hydraenidae	A	A	A	A	B
Hydrophilidae	B	B/C	B/C	C	B
Liodessus	B/C	B/C	B/C	C	B
Onychohydrus	B/C	B/C	B/C	B/C/D	B
Podaena	A	A	A	A	B
Ptilodactylidae	A	A	A	A	B
Rhantus	B/C/D	B/C	B/C	B/C/D	B
Scirtidae	A	A	A	A	B
Staphylinidae					
Neuroptera					
Kempynus	A	A	A		
Diptera					
Anthomyiidae	B/C	B/D	A/D	A/D	
Aphrophila	B	B	B	B	B
Austrosimulium	A	A/C	A	A/B	A/D
Calopsectra					
Ceratopogonidae	A/B/C	A/B/C	A/B/C	A/B	
Chironomidae	C	D	B/D	C	C
Chironomus (Chironomini)	D	B/D	D	D	

Table A3.1, continued

Taxon	Deposited sediment	Nutrients	Oxygen	Temp	Metals
Chironomini (excl. *Chironomus*)	C	B/D	B	B	
Diamesinae	B	A/D	A	A	
Corynoneura	B	A/C	A/C	A/C	
Cryptochironomus		A	B	B	
Culex	C/D	B/C/D	D	D	
Culicidae	C/D	B/C	C	C	
Diptera indet.					
Dixidae	C	C	C	C	
Dolichopodidae					
Empididae	C	C	A/C	A/B/C	B
Ephydridae Eriopterini	C/D C	B/D B	D A/B/C	C A/B	
Harrisius	A	A	A	A	
Hexatomini	B	A	A	A	
Limnophora	B/C	A/C/D	A/C	A/C	
Limonia	A/C	A/C/D	A/C	A	
Lobodiamesa	B	A/D	A	A	
Maoridiamesa	B	A/D	A	A	
Mischoderus	B/C	A	A	A/C	
Molophilus	B/C	A/C	A/C	A	
Muscidae	B/C	B/D	A/C	A/C	C
Nannochorista	A/B/C	A/B	A	B	
Neocurupira	A	A	A	A	A
Neolimnia	C	A/D	A	A	
Nothodixa	B/C/D	B/C	C	C	
Diptera					
Orthocladiinae	B	D	B	B	C/D
Parochlus					
Paradixa	C	B/C	B/C	B/C	
Paralimnophila	B	A	A	A	
Paucispinigera	B	B	B	B	
Pelecorhyncidae	B	A	A	A	
Peritheates	A	A	A	A	
Podonominae					
Polypedilum	B	B	B	B	
Psychodidae	B/D	B	B/D	B/C	
Scatella	C	C	C	C	
Sciomyzidae	C/D	C	C/D	C	
Stratiomyidae	D	B/D	B/C/D	C	
Syrphidae	D	D	D	C/D	
Tabanidae	C	A/D	A/C/D	A/C	
Tanypodinae	B	C	B	B	
Tanytarsini	C	C	B/C	B/C	
Tanytarsus	C/D	C/D	B	B	
Thaumaleidae	A	A	A	A	
Tipulidae	B/C	B/C/D	A/C	A	
Zelandotipula	B	A	A	A	
Trichoptera					
Alloecentrella	A	A	A/B	A	
Aoteapsyche	B	D	B	B	B/C
Beraeoptera	A	A/C	A	A	
Confluens	A	A	A/B	A	
Conuxia	A	A	A/B	A	
Costachorema	A	A	A/B	A/B	A
Cryptobiosella	A	A	A	A	A
Diplectrona	A	A	A	A	C

Table A3.1, continued

Taxon	Deposited sediment	Nutrients	Oxygen	Temp	Metals
Ecnomina	A/B	A	A	A	
Edpercivalia	A	A	A	A	A
Ecnominidae	B	A	A	A	
Helicopsyche	A	A/B	A	A	D
Hudsonema	B	A/B/C	B	B/C	B
Hydrobiosella	B	A	A/B	A/B	B
Hydrobiosis	B	A/B/C	A	B	B
Hydrobiosis (excl. Hydrobiosis below)	B	A	A	A	
Hydrobiosis parumbripennis	B	A/C	B	B	
Hydrobiosis umbripennis/copis/budgei	B	A/C	B	B	
Hydrobiosis styx/torrentis	A	A	A	A	
Hydrochorema	A	A	A	A	B
Kokiria	A/C	A	A	A	D
Neurochorema	A	A	A	A	B
Oecetis	B/C	A	A/B/C	A	
Oeconesidae	A	A	A/B	A	B
Olinga	A	B	A	A	B
Orthopsyche	A	A	A	A	D
Oxyethira	C	D	C	C	D
Paroxyethira	C	D	B/C/D	C	D
Philorheithrus	A	A	A	A	B
Plectrocnemia	B	A	A	A	C
Polyplectropus	B/C	A	B	B	C
Psilochorema	A/B	A	A/B	A/B	D
Psilochorema mimicum	C	A	A	B	
Psilochorema leptoharpax	B	A	A	A	
Pycnocentrella	A	A/B	A/B	A	B
Pycnocentria	B	B	B	B	D
Pycnocentria evecta	B	A/C	A/B	B	
Pycnocentrodes	B	B	B	B	D
Rakiura	A	A	A	A	D
Synchorema	A	A	A	A	
Tiphobiosis	A/B	A	A/B	A/B	C
Triplectides	C	A/C	C	C	C
Triplectidina	B/C	A	A/B/C	A/C	C
Zelandoptila	B	A	A	A	
Zelolessica	A	B	A/B	A	
Lepidoptera					
Hygraula	D	A/B/D	B	B	
Collembola					
ACARINA					
ARACHNIDA					
Dolomedes					
MOLLUSCA					
Gundlachia = Ferrissia	A/C	B	B/C	B/C	
Glyptophysa = Physastra	C/D	D	D	C/D	
Gyraulus	C	D	D	B/C/D	A
Echyridella	D	A/B/D	B/C/D	B	
Latia	A/C	B	A/C	A/C	
Lymnaeidae	C/D	C	D	D	
Melanopsis	B/D	A/D	B/D	B/D	
Physa = Physella	D	D	D	D	B
Potamopyrgus	D	D	D	D	B
Sphaeriidae	C/D	B	C	C/D	B

Table A3.2. Full references for 18 published studies and short references for 8 unpublished studies from which datasets were collated for stressor-specific metric development.

Study	Citation
Blakemore 2012	Blakemore 2012 – unpublished BSc Honours thesis, University of Otago
Burdon et al. 2013	Burdon FJ, McIntosh AR, Harding JS 2013. Habitat loss drives threshold response of benthic invertebrate communities to deposited sediment in agricultural streams. Ecological Applications 23: 1036-1047.
Clapcott 2017, unpubl	Clapcott 2017 – unpublished data collected for MfE sediment project
Harding & Jellyman 2015	Harding JS, Jellyman PG 2015. Earthquakes, catastrophic sediment additions and the response of urban stream communities. New Zealand Journal of Marine and Freshwater Research 49: 346-355.
Jellyman & Harding 2011	Jellyman & Harding 2011 – data from Hurunui report
Lange et al. 2014	Lange K, Townsend CR, Matthaei CD 2014. Can biological traits of stream invertebrates help disentangle the effects of multiple stressors in an agricultural catchment? Freshwater Biology 59: 2431-2446.
Magbanua et al. 2010	Magbanua FS, Townsend CR, Blackwell GL, Phillips N, Matthaei CD 2010. Responses of stream macroinvertebrates and ecosystem function to conventional, integrated and organic farming. Journal of Applied Ecology 47: 1014-1025.
Quinn & Hickey 1990	Quinn JM, Hickey CW 1990. Characterisation and classification of benthic invertebrate communities in 88 New Zealand rivers in relation to environmental factors. New Zealand Journal of Marine and Freshwater Research 24: 387-409.
Quinn & Hickey 1993	Quinn JM, Hickey CW 1993. Effects of sewage waste stabilization lagoon effluent on stream invertebrates. Journal of Aquatic Ecosystem Health 2: 205-219.
Ramezani et al. 2016	Ramezani J, Akbaripasand A, Closs GP, Matthaei CD 2016. In-stream water quality, invertebrate and fish community health across a gradient of dairy farming prevalence in a New Zealand river catchment. Limnologica - Ecology and Management of Inland Waters 61: 14-28
Storey et al. 2009	Storey R, Parkyn S, Smith B, Croker G, Franklin P 2009 – Effects of development on zero-order streams in the Waikato region. Environment Waikato Technical Report 2009/22
Townsend 2008, survey	Townsend CR, Uhlmann SS, Matthaei CD 2008. Individual and combined responses of stream ecosystems to multiple stressors. Journal of Applied Ecology 45: 1810-1819.
Wagenhoff et al. 2011	Wagenhoff A, Townsend CR, Phillips N, Matthaei CD 2011. Subsidy-stress and multiple-stressor effects along gradients of deposited fine sediment and dissolved nutrients in a regional set of streams and rivers. Freshwater Biology 56: 1916-1936.
Wagenhoff et al. 2017	Wagenhoff A, Liess A, Pastor A, Clapcott JE, Goodwin EO, Young RG 2017. Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams. Freshwater Science: 36: 178-194.
Collier & Smith 2005	Collier KJ, Smith BJ 2005. Effects of progressive catchment harvesting on stream invertebrates in two contrasting regions of New Zealand's North Island. Marine and Freshwater Research 56: 57-68.
Eivers 2006	Eivers 2006 – unpublished MSc thesis, University of Canterbury
Graham & Quinn Whatawhata, unpubl.	Graham E, Quinn J – unpublished Whatawhata data 1995-2013
Holmes 2008	Holmes 2008 – unpublished MSc thesis, University of Otago
Matthaei et al. 2006	Matthaei CD, Weller F, Kelly DW, Townsend CR 2006. Impacts of fine sediment addition to tussock, pasture, dairy and deer farming streams in New Zealand. Freshwater Biology 51: 2154-2172.
Reid et al. 2010	Reid DJ, Quinn JM, Wright-Stow AE 2010. Responses of stream macroinvertebrate communities to progressive forest harvesting: Influences of harvest intensity, stream size and riparian buffers. Forest Ecology and Management 260: 1804-1815.

Table A3.2, continued

Study	Citation
Townsend 2008, experiment	Townsend CR, Uhlmann SS, Matthaei CD 2008. Individual and combined responses of stream ecosystems to multiple stressors. Journal of Applied Ecology 45: 1810-1819.
HBRC 2016	HBRC 2016 – unpublished Hawke's Bay Regional Council data from Waihi dam failure
Matthaei et al. 2010	Matthaei CD, Piggott JJ, Townsend CR 2010. Multiple stressors in agricultural streams: interactions among sediment addition, nutrient enrichment and water abstraction. Journal of Applied Ecology 47: 639-649.
Piggott et al. 2012	Piggott JJ, Lange K, Townsend CR, Matthaei CD 2012. Multiple stressors in agricultural streams: a mesocosm study of interactions among raised water temperature, sediment addition and nutrient enrichment. Plos One 7: e49873.
Piggott et al. 2015	Piggott JJ, Townsend CR, Matthaei CD 2015. Climate warming and agricultural stressors interact to determine stream macroinvertebrate community dynamics. Global Change Biology 21: 1887-1906.
Wagenhoff et al. 2012	Wagenhoff A, Townsend CR, Matthaei CD 2012. Macroinvertebrate responses along broad stressor gradients of deposited fine sediment and dissolved nutrients: a stream mesocosm experiment. Journal of Applied Ecology 49: 892-902

Table A3.3. Summary of research datasets collated for stressor-specific metric development. See text for description of variables.

Study name	Study approach	Invertebrate data	No. of sites / exp. units	No. of samples	Deposited sediment variables	Nutrient variables	Periphyton variables
Blakemore 2012	survey	relative abundance	43	43	sedcover_instream, SIS	TN, NO2_NO3, NH4, DIN, TP, DRP	chl-*a*
Burdon et al. 2013	survey	density	30	30	sedcover_instream, SIS	NO3, DRP	
Clapcott 2017, unpubl	survey	density	16	16	sedcover_bankside, SIS		
Harding & Jellyman 2015	survey	density	16	16	sedcover_instream, sedcover_bankside		
Jellyman & Harding 2011	survey	coded abundance	42	42		TN, NO3, NO2, NH4, DIN, DRP	
Lange et al. 2014	survey	relative abundance	43	43	sedcover_instream, SIS,	TN, NO2_NO3, NH4, DIN, TP, DRP	chl-*a*
Magbanua et al. 2010	survey	density	30	30	sedcover_instream, SIS,	TN, NO3, NH4, TP, DRP	
Quinn & Hickey 1990	survey	density	88	88	sedcover_instream	TKN, NO3, NH4, TP, DRP	chl-*a*, AFDM
Quinn & Hickey 1993	survey	density	11	11		NH4, DIN, DRP	
Ramezani et al. 2016	survey	relative abundance	36	36	SIS	TN, NO2_NO3, NH4, DIN, TP, DRP	
Storey et al. 2009	survey	relative abundance	46	46	wolman		
Townsend 2008, survey	survey	density	32	32	sedcover_instream, SIS	NO2_NO3, DIN, DRP	chl-*a*, AFDM
Wagenhoff et al. 2011	survey	relative abundance	43	43	sedcover_instream, SIS	NO2_NO3, DIN, DRP	chl-*a*, AFDM
Wagenhoff et al. 2017	survey	relative abundance	58	58	wolman, sedcover_bankside	TN, NO3, NH4, DIN, TP, DRP	chl-*a*, peri_cover_fils
Collier & Smith 2005	field experiment	density	8	68	wolman		
Eivers 2006	field experiment	coded abundance	51	51	sedcover_instream, SIS		chl-*a*
Graham & Quinn Whatawhata, unpubl.	field experiment	density	9	386	wolman, SIS	TN, TKN, NO3, NH4, TP, DRP	chl-*a*, AFDM, peri_cover_total
Holmes 2008	field experiment	density	9	18	sedcover_instream	NO2_NO3, NH4, DIN, DRP	chl-*a*
Matthaei et al. 2006	field experiment	density	24	48	sedcover_instream	NO2_NO3, NH4, DIN, DRP	
Reid et al. 2010	field experiment	density	20	424	wolman		AFDM
Townsend 2008, exp.	field experiment	density	18	18	sedcover_instream	NO2_NO3, NH4, DIN, DRP	chl-*a*
HBRC 2016	field experiment	density	8	22	wolman, SIS	TN, DIN, TP	
Matthaei et al. 2010	mesocosm experiment	density (on tile)	18	18	sedcover_instream	NO3, NH4, DRP	chl-*a*, peri_cover_total
Piggott et al. 2012	mesocosm experiment	density (on tile)	18	18	sedcover_instream	NO3, NH4, DRP	chl-*a*
Piggott et al. 2015	mesocosm experiment	density	128	128	sedcover_instream	NO2_NO3, NH4, DIN, DRP	chl-*a*
Wagenhoff et al. 2012	mesocosm experiment	density	128	128	sedcover_instream, 'SIS'	DIN, DRP	chl-*a*

Figure A3.1. Histograms of the raw and transformed macroinvertebrate relative abundance data for the 'sediment cover' macroinvertebrate dataset ('SIS' macroinvertebrate dataset not shown). RelAbd_raw = relative abundance data (proportion), RelAbd_lograw = natural log-transformation of the proportional data, RelAbd_logpct = natural log-transformation of the percentage data, RelAbd_logit = logit-transformation of the proportional data.

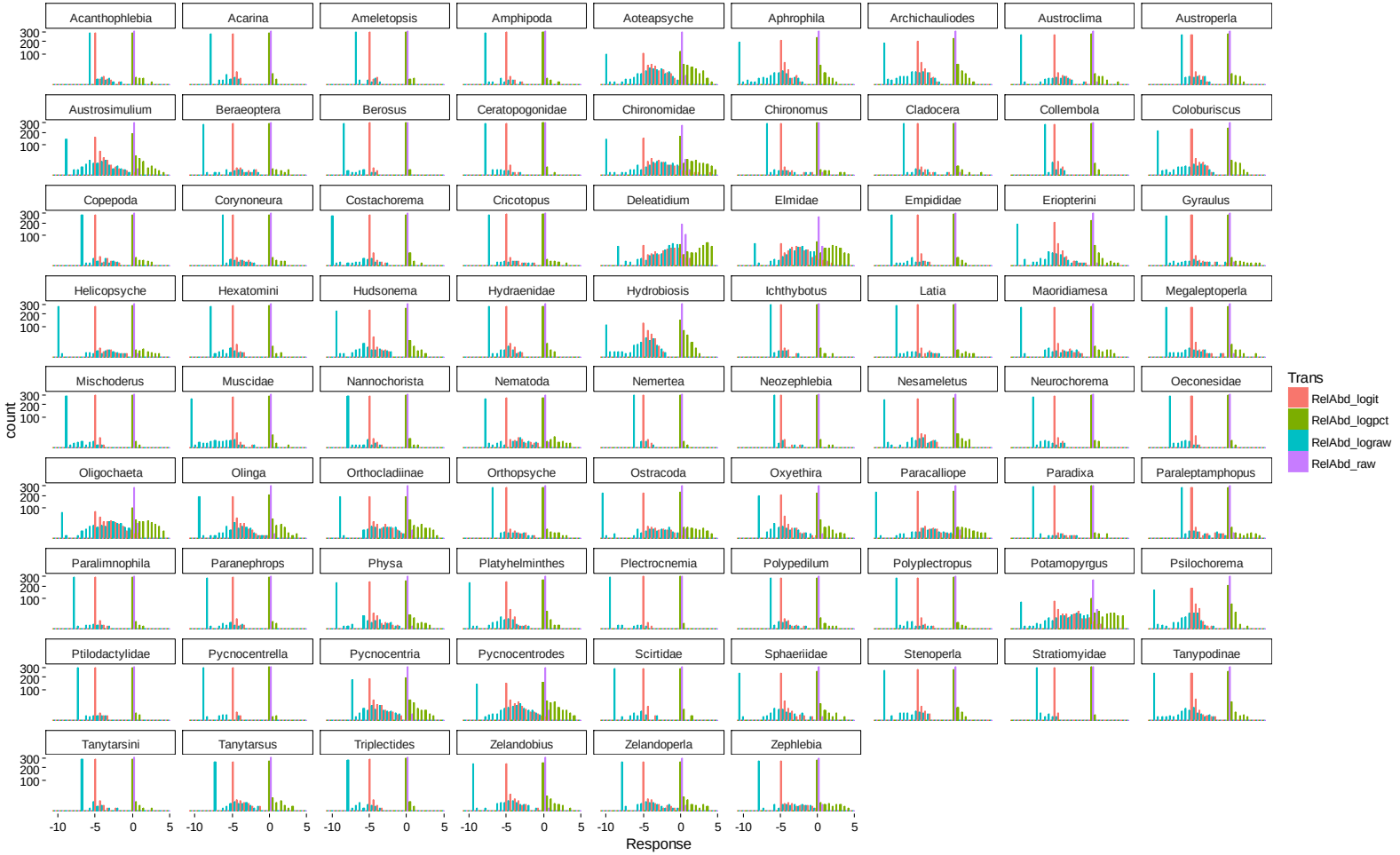
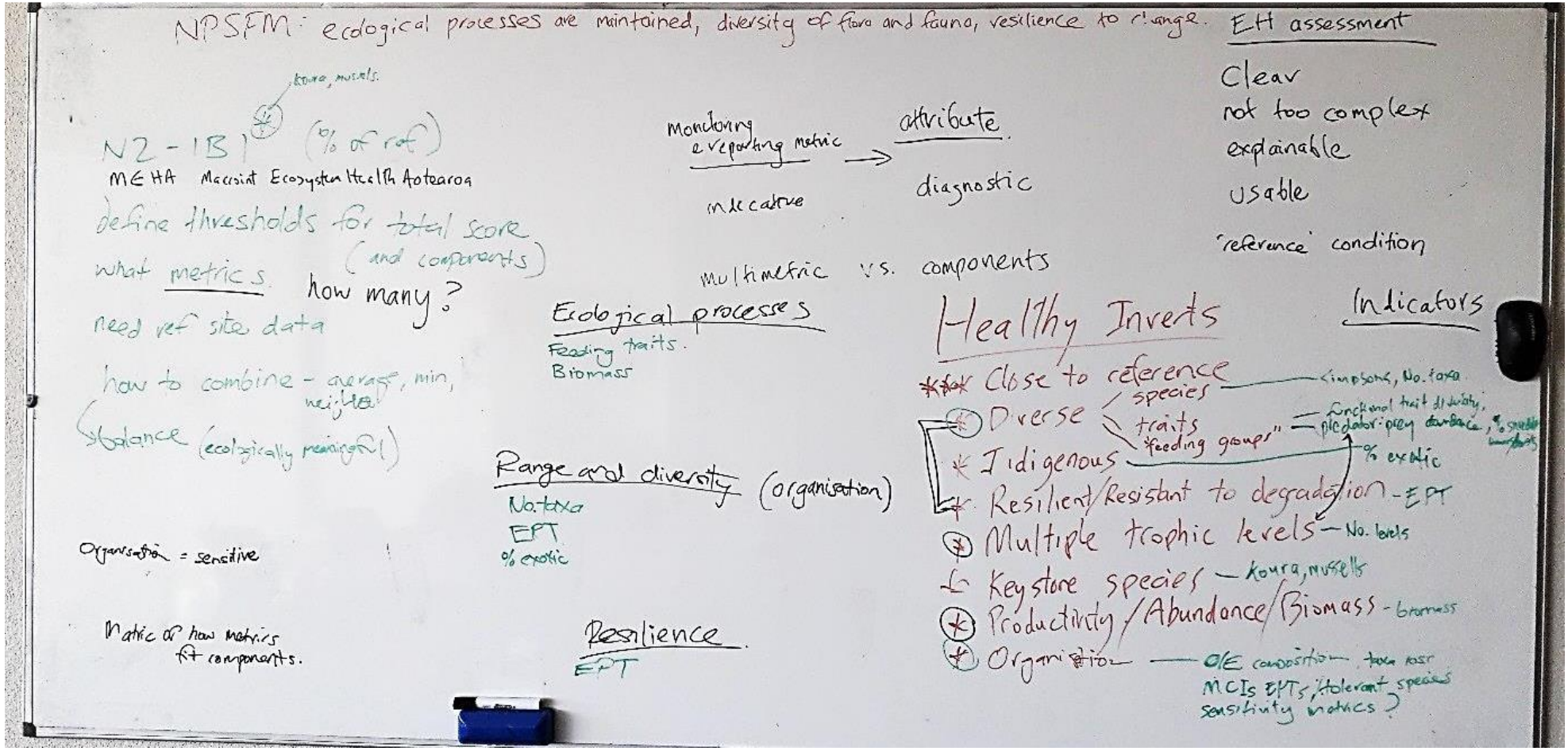


Figure A3.2. Image of the whiteboard summarising team discussion from Workshop 2.



Appendix 4. Linking metrics to stressors: dataset compilation

Collation of existing national and research datasets is described in Sections 2.1.1 and 2.1.2, respectively. However, only subsets of these large datasets were used for linking metrics to stressors.

National macroinvertebrate-stressor dataset

The national macroinvertebrate dataset contains SoE data provided by regional and unitary councils as well as data collected by NIWA from National River Water Quality Network (NRWQN) sites[10] typically collected on an annual basis (see details in Sections 2.1.1). This dataset was matched with stressor data retrieved from three separate datasets.

1) A dataset compiled for a parallel MfE project (Depree et al. 2017) consisting of deposited and suspended fine sediment data.

   Deposited sediment measures were:
   - bankside visual assessment of sediment cover within the Rapid Habitat Assessment (RHA) protocol
   - bankside visual assessment of sediment cover (% cover bankside, SAM1)
   - instream visual assessment of sediment cover (% cover instream, SAM2)
   - Wolman pebble count (% fines, SAM3)
   - suspendable inorganic sediment (SIS, SAM4)
   - suspendable benthic sediment volume (SBSV, SAM4)
   - shuffle test score (SAM5).

   Suspended sediment measures included:
   - total suspended solids (TSS)
   - turbidity
   - visual clarity.

The frequency of deposited sediment assessments at a single site, in particular, generally varies largely. In some instances, assessments were made monthly, but in most cases deposited sediment assessments were made annually, or had been done only once for a site.

2) Water quality data collected at SoE monitoring sites (typically monthly), was retrieved from the LAWA (Land, Air and Water Aotearoa) website (downloaded 5 May 2017). The following water quality measures were of interest:
   - ammonium-nitrogen (NH4N)
   - total oxidised nitrogen (NOxN)

[10] Compiled by Martin Unwin, NIWA, Christchurch.

- total nitrogen (TN)
- dissolved reactive phosphorus (DRP)
- total phosphorus (TP).
- turbidity
- visual clarity (black disk)

3) Periphyton data (various measures) also compiled as part of the parallel MfE project (Depree et al. 2017). Data were mostly assessed at SoE or NRWQN sites on a monthly or annual basis. The following periphyton measures were of interest:
    - benthic chlorophyll-*a* (chl-*a*)
    - visual assessment of total periphyton cover.

Due to potentially significant annual variation of sediment, nutrient and periphyton conditions at a single site, we aimed at calculating a median value from all available data collected within the same month and the 12 months prior to macroinvertebrate sampling. However, monthly (or more) observations were not always available. In particular for deposited sediment measures often only a single observation was available within the 12-month time period.

The macroinvertebrate and stressor datasets contained various identifiers with which samples could be matched. Matching deposited and suspended sediment data from the MfE Sediment project with macroinvertebrate data was done first. The matching process first used site name (e.g., 'Makotuku at Raetihi') and sampling date, accounting for possible multiple sediment sampling dates as described above. Inconsistencies in site names e.g., due to use of 'at' or '@', 'Road' or 'Rd', as well as spelling mistakes of rivers, road and place names, may have caused potential matches to be missed. Accordingly, uniform site names were created and 'fuzzy matching' accommodated minor inconsistencies in site names. Exact matches were accepted without further checking whereas fuzzy matches were manually checked using their NZReach IDs. This process increased real matches. For cases where no site name match was found, matching was then performed using regional council site ID (RCSID). Lastly, further matches between macroinvertebrate and sediment data were found via NZReach ID, limited for those NZReach IDs where a single macroinvertebrate site existed. There were multiple occasions where sample sites were in close proximity, typically upstream and downstream of sewage treatment plants, resulting in very similar uniform site names and both having the same NZReach ID. These cases were matched manually before matching by NZReach.

After matching of macroinvertebrate with sediment data, macroinvertebrate data were matched with LAWA data, first by matching site name and date using the same fuzzy matching approach. Secondly, further matches were found by comparing regional council site ID (RCSID) and LAWA ID. Thirdly, further matches were found using NZREACH ID for those cases where only a single site was sampled within a NZReach

ID. Finally, periphyton data were matched with macroinvertebrate data using site name and date, RCSID and NZReach ID as above.

Merge with research data

Research data combining macroinvertebrate and stressor data was compiled from a total 26 studies (see details in Section 2.1.2). This research dataset predominately contains stressor data from a single observation taken on or close to the day of sampling macroinvertebrates. The same subset of this large research dataset was used as that used for stressor-specific metric development. More details on the research data selection process and rational are described in Section 3.4.2. Finally, selected research data were merged with the national macroinvertebrate-stressor dataset.

Data checking

For the majority of sample sites NZReach ID was known, and predictions of environmental variables were retrieved from various existing databases. For example, the percentage of intensive pastoral land use in the catchment (T2PastoralHeavy) and the percentage of native vegetation cover (T1NativeVeg) calculated from the Land Use Cover Data Base 3 (LCDB3) were retrieved and bivariate scatterplots produced for various macroinvertebrate metrics. These scatterplots were investigated for unusual values among the stressor attributes. We also checked data distributions of a set of macroinvertebrate metrics and compared summary statistics of these metrics between the national and research datasets. We also visually investigated if summary statistics of the macroinvertebrate metrics differed according to the sampling method (quantitative vs. semi-quantitative). Overall, the data looked fine.

Appendix 5. Linking metrics to stressors: supplementary material from statistical analyses in Section 5.

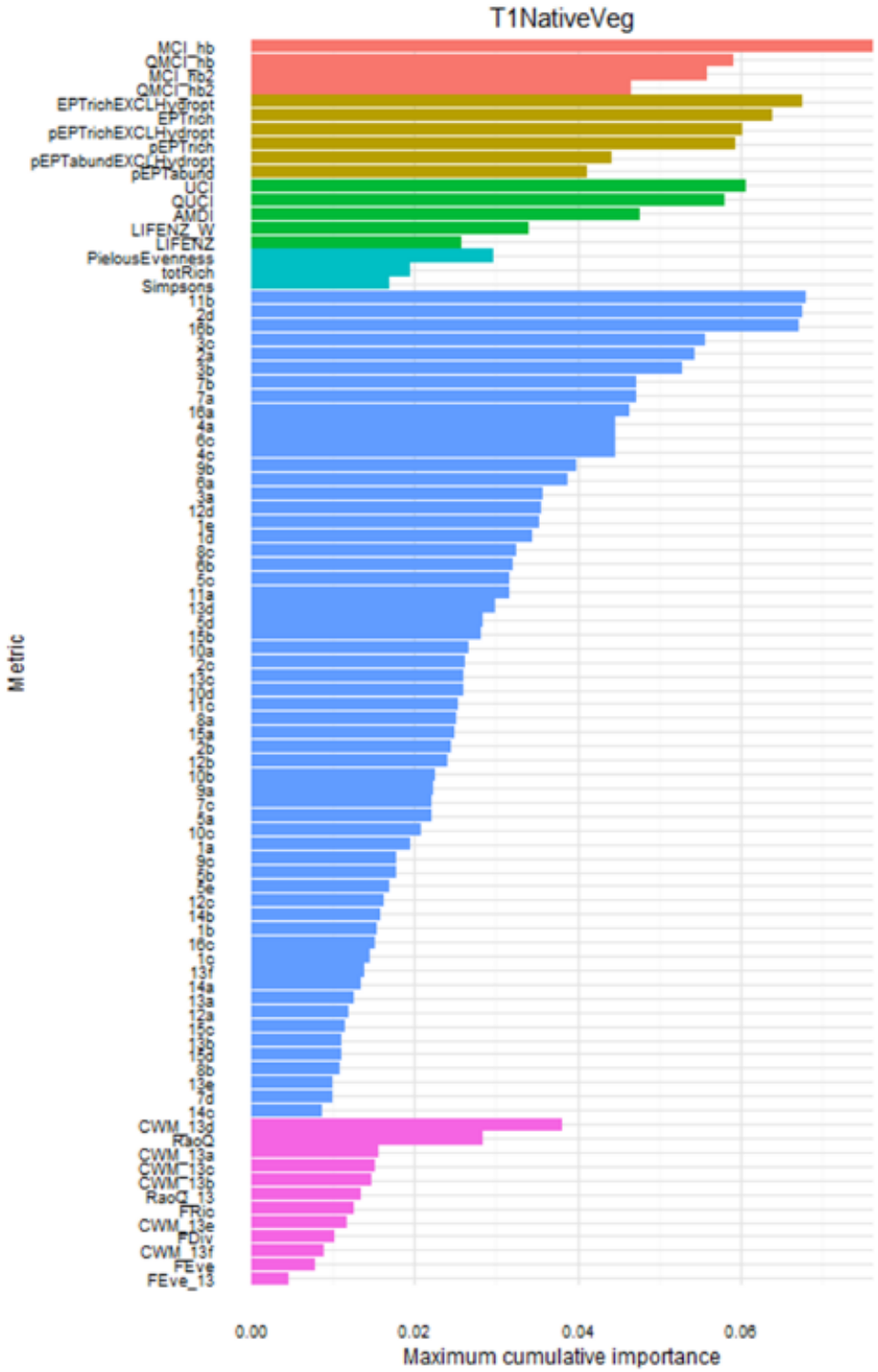


Figure A5.1. Maximum cumulative importance for each of the 90 macroinvertebrate metrics across the native vegetation and pastoral heavy land cover gradients shown in descending order for each metric group (shown by different colours).

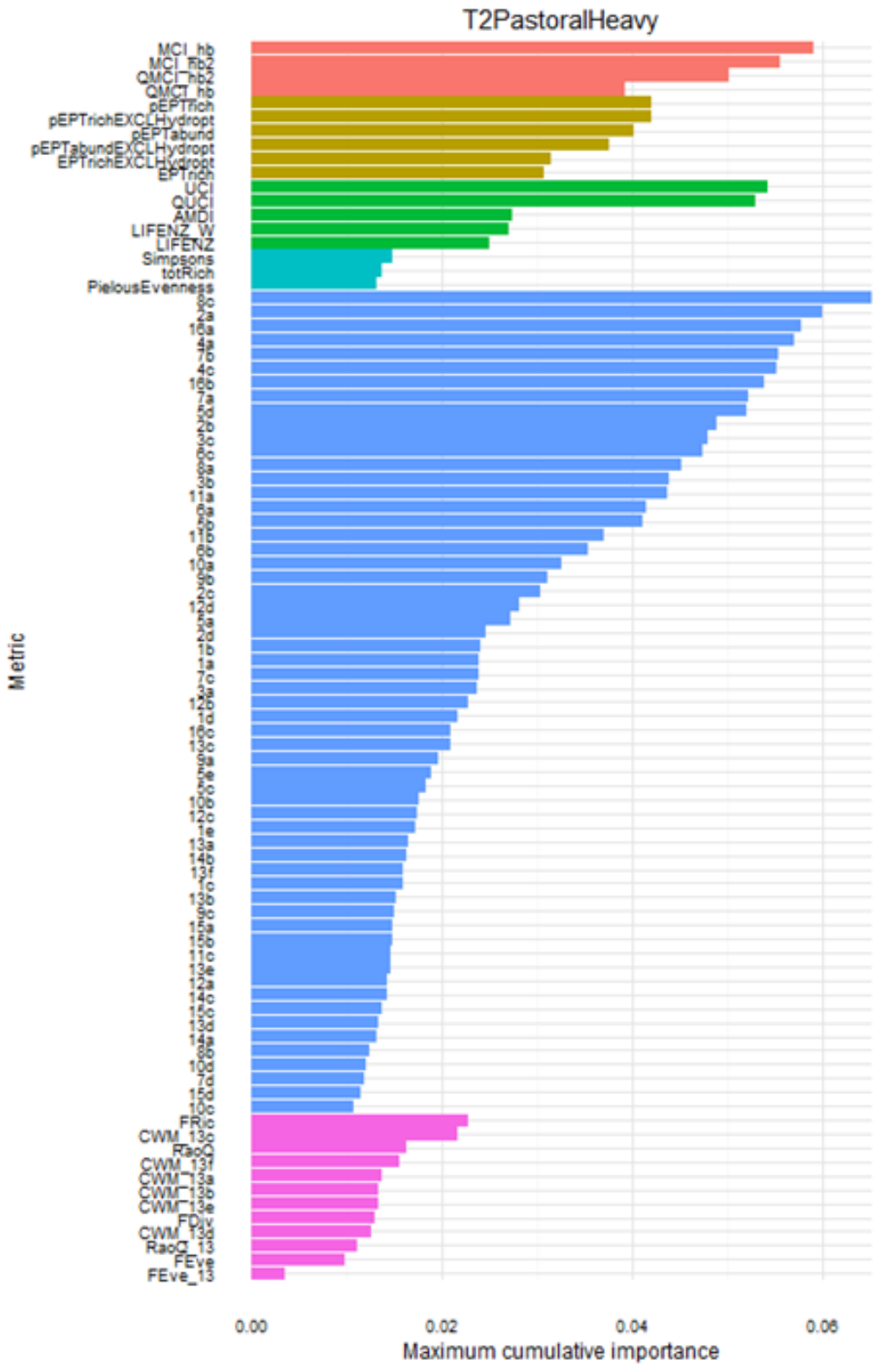


Figure A5.1, continued.

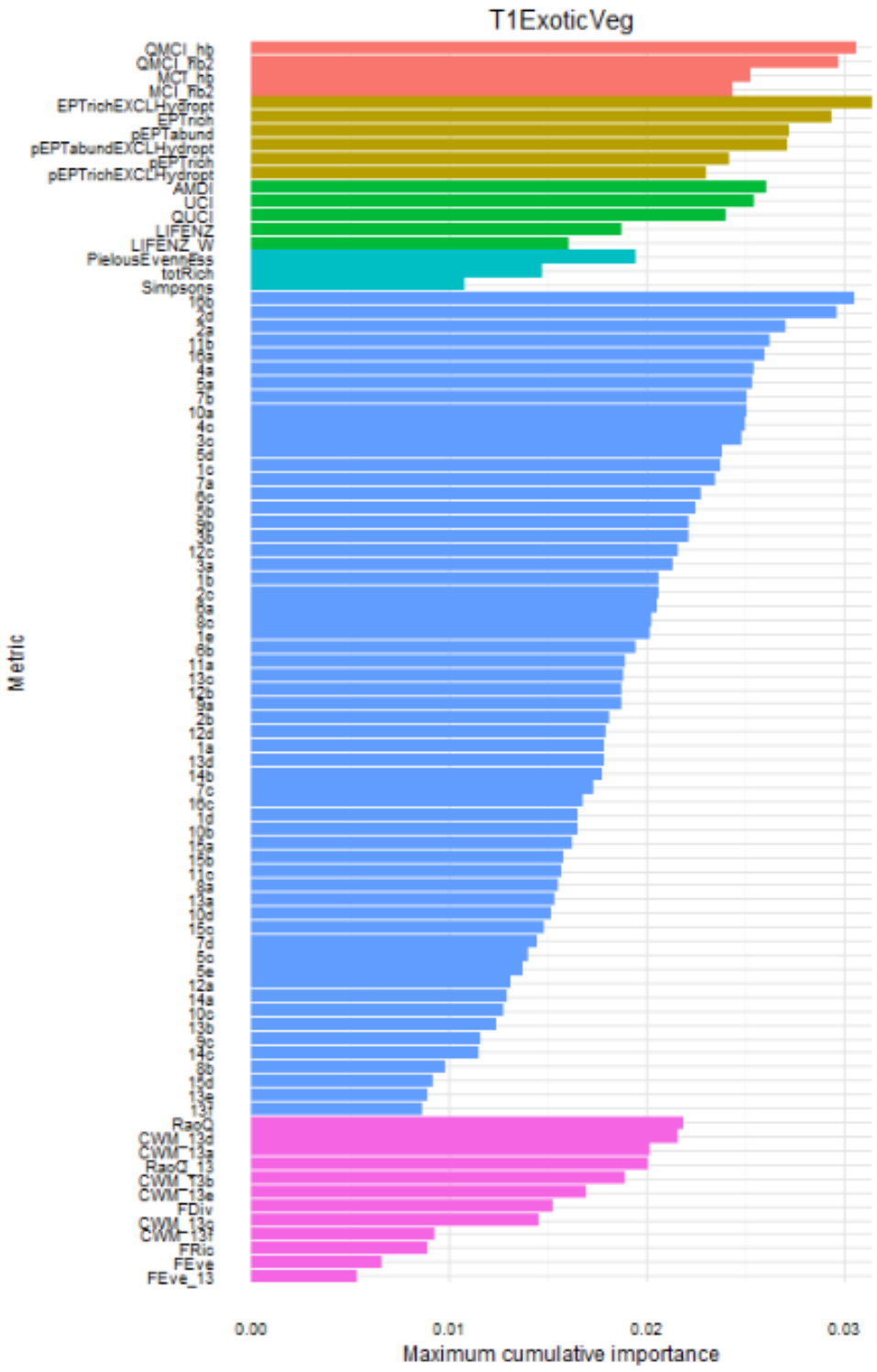


Figure A5.2. Maximum cumulative importance for each of the 90 macroinvertebrate metrics across the exotic vegetation and urban land cover gradients shown in descending order for each metric group.

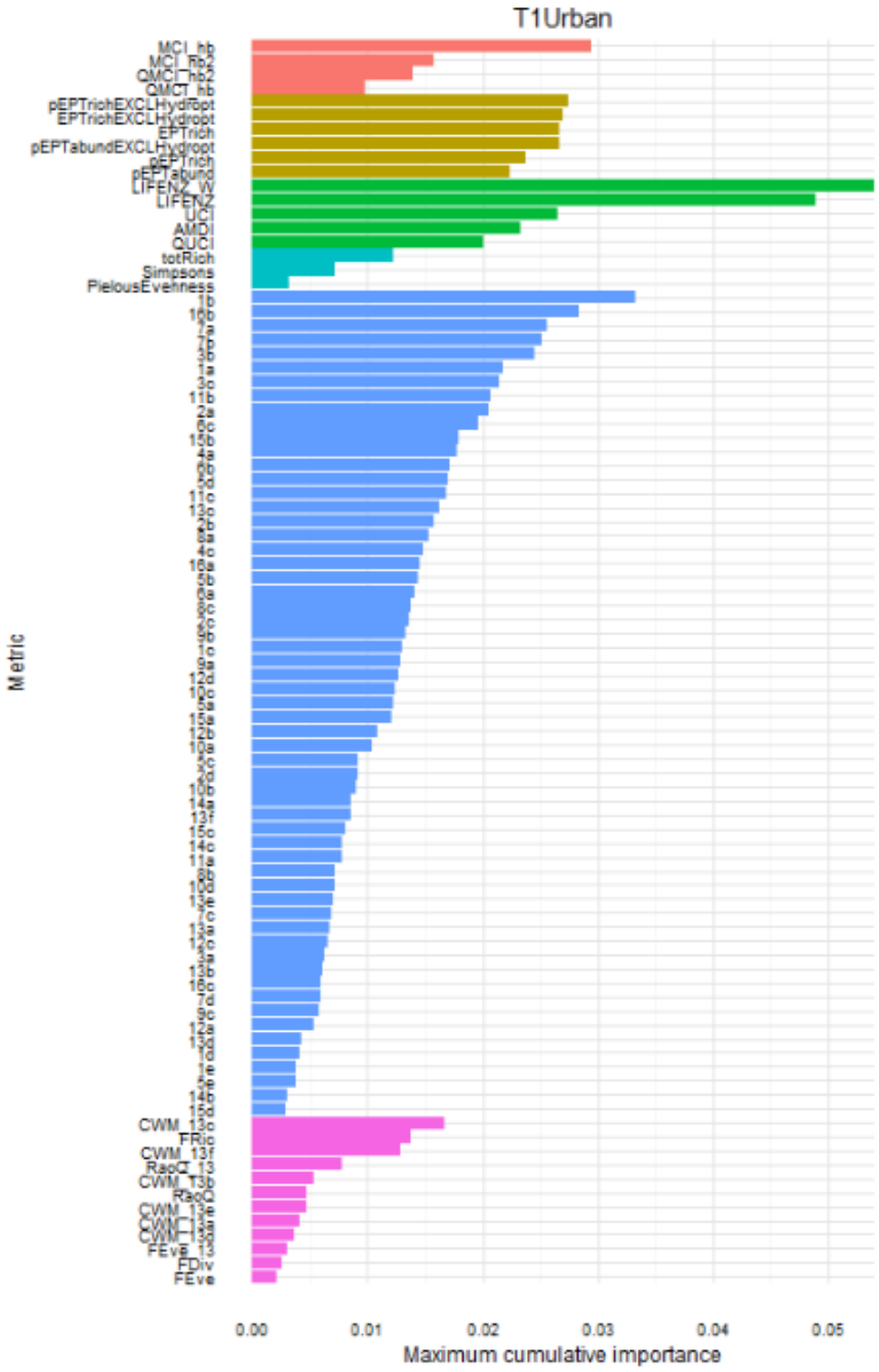


Figure A5.2, continued.

Table A5.1. Ranks of relative importance of predictors from random forest models of macroinvertebrate metrics in response to catchment-scale stressors and environmental descriptors. The R^2 values of the random forest models are also shown.*excluding Hydoptilidae

Metric	MCI_hb	QMCI_hb	MCI_hb2	QMCI_hb2	EPTrich	pEPTrich	pEPTabund	EPTrich*	pEPTrich*	pEPTabund*	LIFENZ	LIFENZ_W	AMDI	UCI	QUCI	totRich	Evenness	Simpsons
R2	0.76	0.60	0.81	0.70	0.65	0.67	0.64	0.68	0.71	0.68	0.81	0.76	0.62	0.75	0.67	0.40	0.39	0.30
T1NativeVeg	1	1	4	5	2	2	5	2	2	5	15	13	2	2	3	9	2	7
T1ExoticVeg	16	8	15	9	11	15	10	11	17	12	18	18	14	16	13	14	12	17
T2PastoralHeavy	4	6	5	4	9	5	6	9	6	7	16	15	11	6	4	18	17	9
T1Urban	14	19	17	18	16	16	16	16	13	13	6	5	16	14	17	19	19	19
maxrateToQ50	18	17	18	17	18	18	18	18	18	18	17	17	19	18	18	5	16	1
DSDIST2COA	17	11	14	12	15	13	13	15	14	15	12	12	15	15	9	12	4	18
ELEVATION	2	5	7	7	1	3	2	1	3	2	2	1	1	5	2	1	15	5
FRE3	13	12	11	10	14	14	11	12	15	10	9	10	9	12	10	10	1	12
ORDER_	19	18	19	19	19	19	19	19	19	19	19	19	18	19	19	13	18	14
SEGFLOWSTA	6	9	3	6	4	4	4	4	4	4	3	4	5	1	6	11	7	11
SEGJANAIRT	3	2	1	1	3	1	1	3	1	1	1	2	3	7	1	6	10	4
SEGMINTNOR	12	14	13	15	10	10	9	10	10	9	13	16	8	11	14	3	3	13
SEGRIPSHAD	15	15	16	16	13	17	15	13	16	16	14	7	4	17	15	2	11	3
SpecMALF	10	16	9	14	7	8	17	8	7	17	4	3	10	8	11	17	8	16
SpecMeanF	7	13	6	11	6	7	12	6	8	11	8	9	7	4	12	16	6	15
USAVGSLOPE	5	4	2	2	8	9	3	7	9	3	10	14	12	3	5	8	14	6
USCALCIUM	11	3	12	3	12	12	8	14	11	8	11	8	17	13	7	7	9	8
USHARDNESS	9	10	10	13	17	11	14	17	12	14	7	6	13	10	16	15	13	10
USPHOSPHOR	8	7	8	8	5	6	7	5	5	6	5	11	6	9	8	4	5	2

Table A5.2. Ranks of relative importance of variables from random forest models of a selection of macroinvertebrate traits in response to catchment-scale stressors and environmental descriptors. The R^2 value for each trait model is also shown.

	1b	1e	2a	2b	2d	3b	3c	4a	4c	5a	5d	7a	7b	8c	10a	11b	12a	13c	14c	15b	15c	16a	16b	CWM_13c
R2	0.58	0.49	0.70	0.63	0.59	0.68	0.69	0.79	0.78	0.70	0.66	0.72	0.74	0.74	0.60	0.63	0.51	0.58	0.37	0.42	0.38	0.71	0.73	0.51
T1NativeVeg	18	3	4	13	1	3	3	8	8	16	12	6	4	10	12	1	18	9	18	3	18	6	2	17
T2PastoralHeavy	12	17	2	4	15	5	4	5	5	10	4	4	3	4	7	7	15	16	13	18	17	3	3	12
T1ExoticVeg	16	14	14	15	10	17	15	15	15	13	14	16	16	16	13	13	17	17	16	16	16	14	13	18
T1Urban	8	19	17	17	19	15	17	17	18	19	18	14	15	18	19	17	19	19	19	14	19	18	15	16
maxrateToQ50	17	15	18	18	17	18	18	18	17	18	16	18	18	17	18	18	3	18	10	12	11	17	18	14
DSDIST2COA	14	16	15	14	7	14	14	14	14	14	13	15	14	14	16	15	9	10	11	7	9	15	16	5
ELEVATION	7	4	1	2	12	2	2	2	2	7	3	2	1	1	6	4	6	2	12	8	8	1	1	1
FRE3	5	6	12	9	3	9	11	10	10	9	6	11	10	9	10	10	5	11	9	5	4	11	14	11
ORDER_	19	18	19	19	18	19	19	19	19	11	19	19	19	19	15	19	16	13	8	19	7	19	19	19
SEGFLOWSTA	2	8	3	3	6	4	5	4	4	3	2	3	5	5	2	5	8	4	4	11	10	4	7	2
SEGJANAIRT	1	1	6	1	9	1	1	3	3	1	5	1	2	3	1	2	1	8	3	4	5	5	4	7
SEGMINTNOR	4	13	11	7	4	10	10	12	12	6	8	9	12	13	9	9	7	7	1	1	2	12	10	6
SEGRIPSHAD	9	12	16	16	2	16	16	16	16	5	17	17	17	15	14	14	11	15	5	13	3	16	11	15
SpecMALF	13	11	10	10	14	11	9	7	6	17	9	13	9	6	3	12	14	5	17	15	13	10	8	9
SpecMeanF	15	5	9	11	5	6	7	6	7	15	10	8	6	8	5	8	10	3	15	9	14	8	6	8
USAVGSLOPE	6	2	5	6	8	7	6	1	1	2	1	7	8	2	4	3	4	1	7	6	12	2	5	3
USCALCIUM	10	10	13	12	13	12	13	13	13	12	15	12	13	12	11	11	12	12	2	10	6	13	12	10
USHARDNESS	11	9	8	8	16	13	12	11	11	8	11	10	11	11	17	16	13	14	14	17	15	9	17	13
USPHOSPHOR	3	7	7	5	11	8	8	9	9	4	7	5	7	7	8	6	2	6	6	2	1	7	9	4

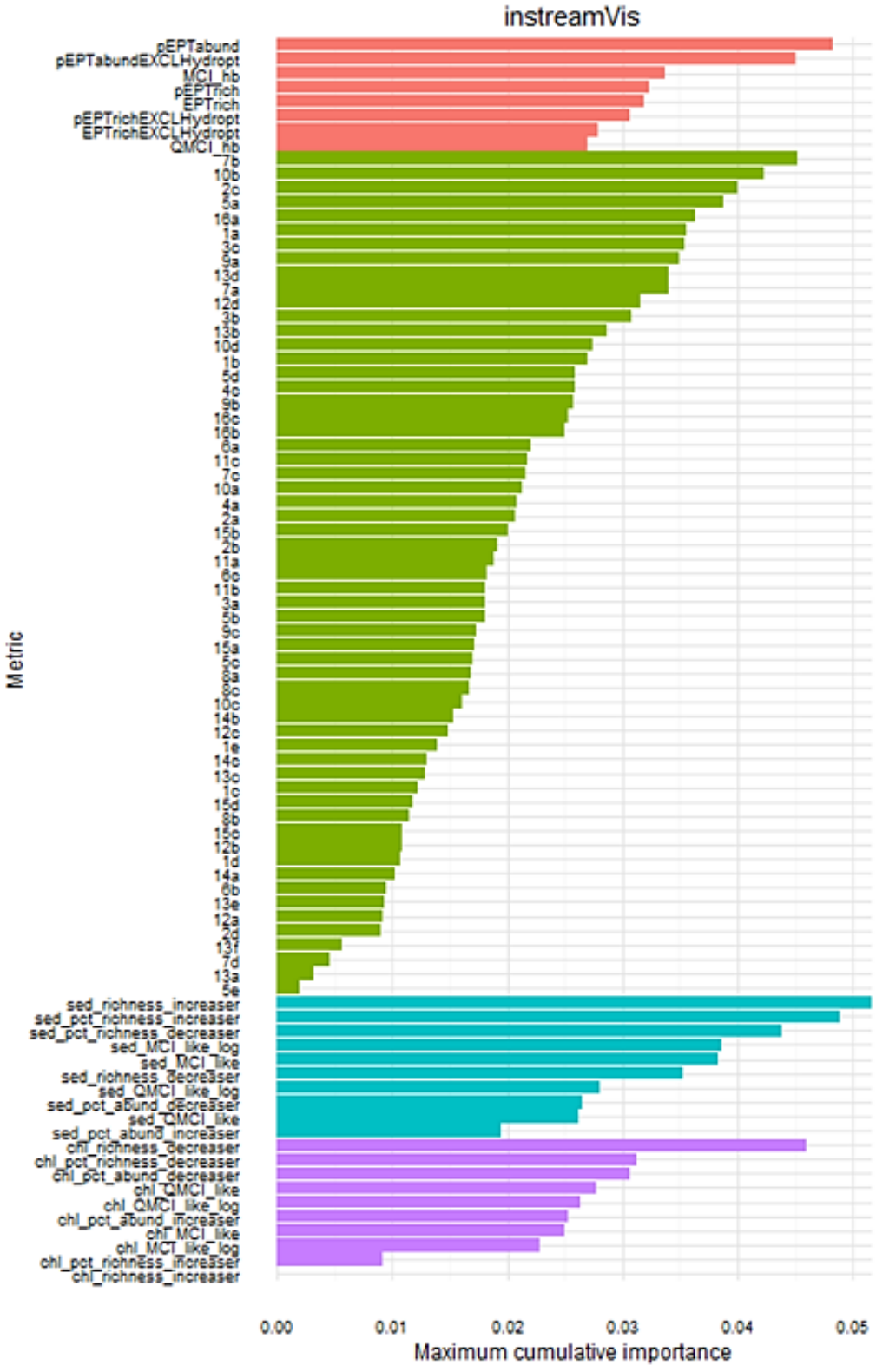


Figure A5.3. Maximum cumulative importance for each of the 88 macroinvertebrate metrics across the sediment (instreamVis) and enrichment (Chla) stressor gradients shown in descending order for each metric group.

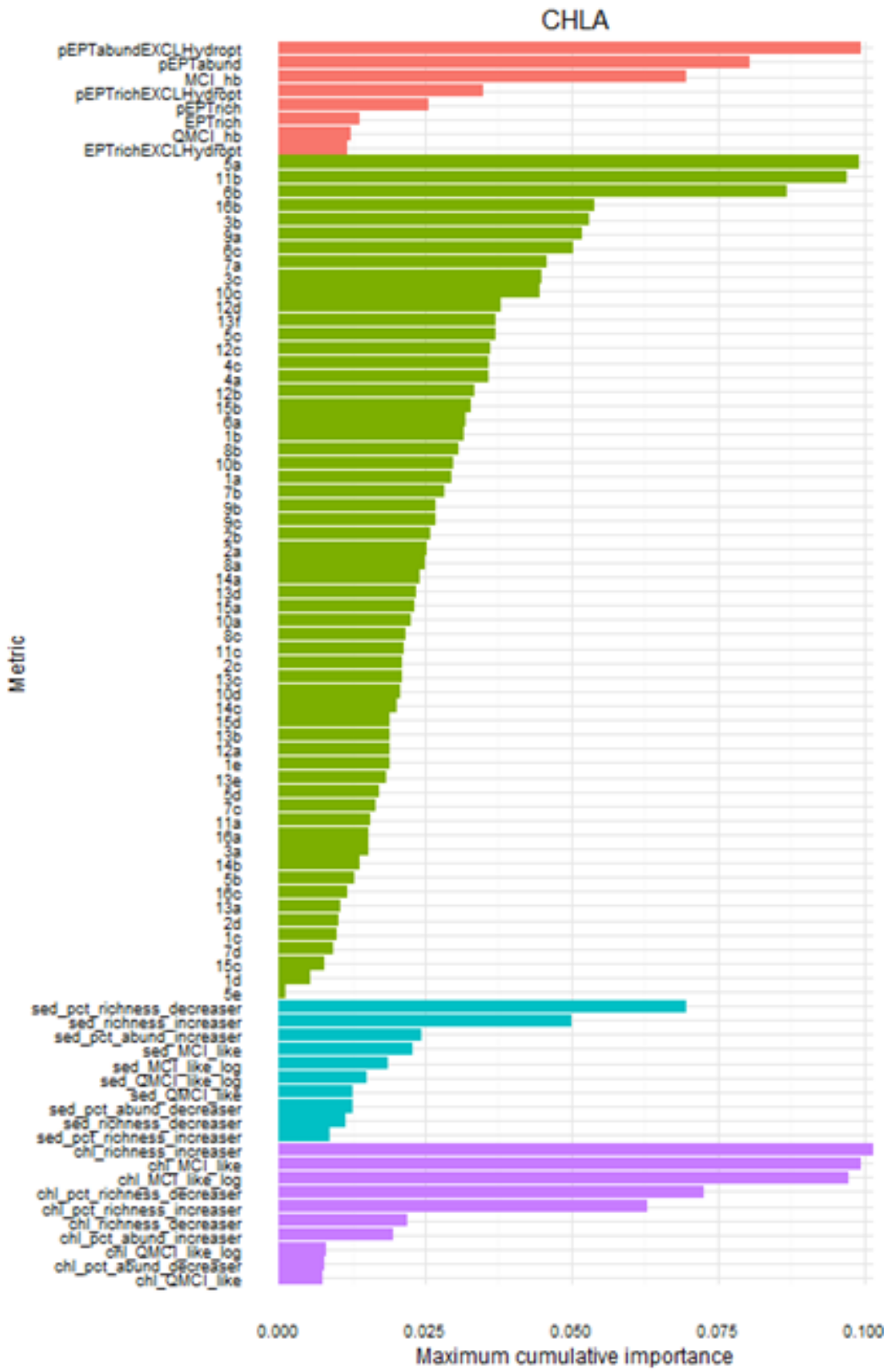


Figure A5.3, continued.

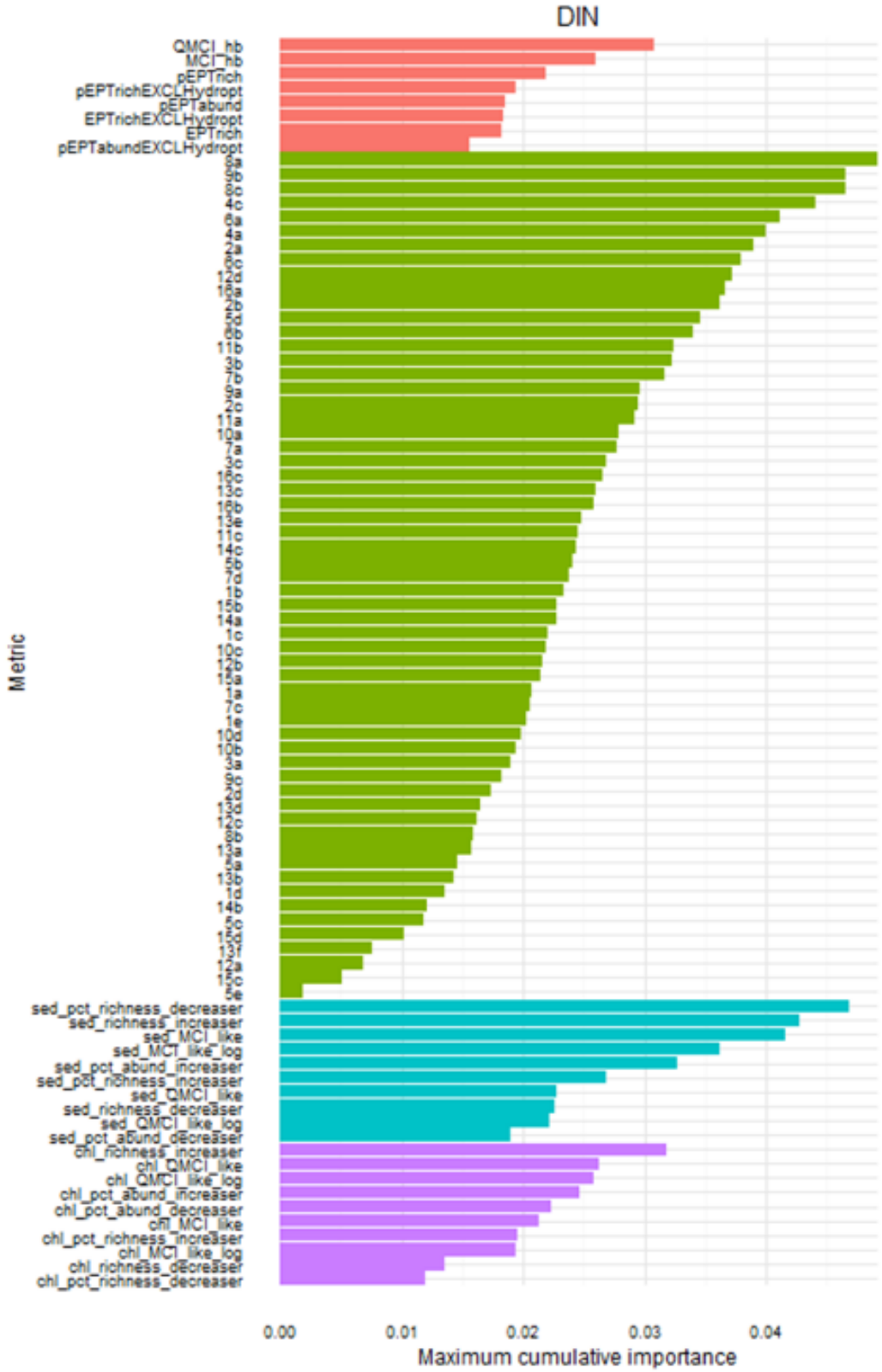


Figure A5.4. Maximum cumulative importance for each of the 88 macroinvertebrate metrics across the nitrogen (DIN) and phosphorus (DRP) stressor gradients shown in descending order for each metric group.

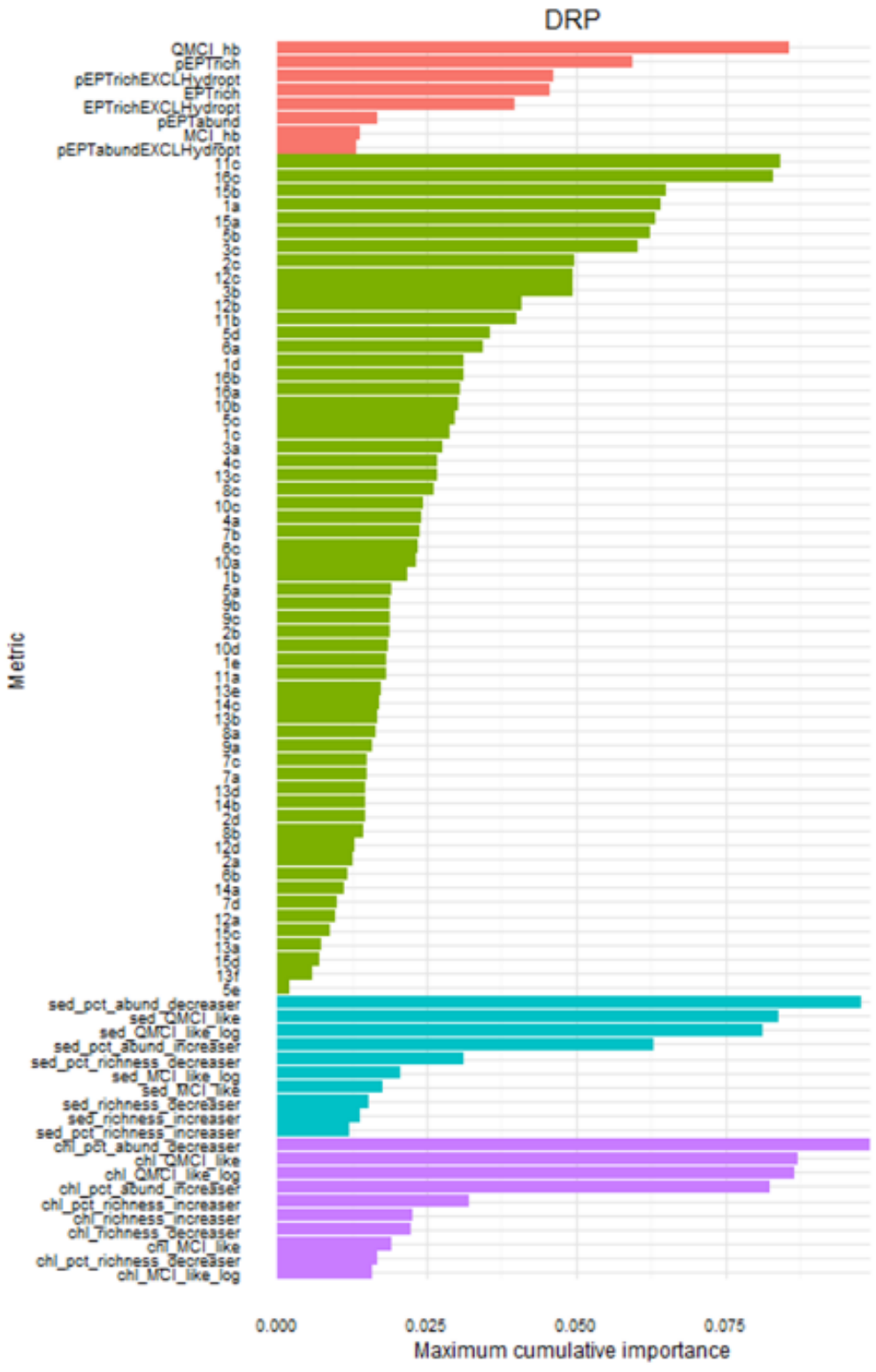


Figure A.5.4, continued.

Table A5.3. Relative importance of predictors by rank for a selection of metrics shown in the GF analysis to be most informative of each stressor. The R^2 values of the random forest models are also shown. *excluding Hydoptilidae.

metric	MCI_hb	pEPTabund*	sed_MCI_like	sed_richness_decreaser	sed_pct_richness_decreaser	sed_richness_increaser	sed_pct_richness_increaser	chl_MCI_like	chl_richness_decreaser	chl_pct_richness_decreaser	chl_richness_increaser	chl_pct_richness_increaser
R^2	0.63	0.49	0.6	0.59	0.67	0.52	0.46	0.57	0.54	0.31	0.41	0.44
CHLA	2	1	13	19	2	4	18	1	14	1	1	1
DIN	12	13	5	13	5	5	8	10	19	10	2	10
DRP	18	17	15	17	8	15	17	14	13	5	6	3
instreamVis	8	3	8	7	6	3	1	7	3	2	19	18
maxrateToQ50	13	7	11	12	16	19	11	8	8	6	10	12
DSDIST2COAST	17	18	18	16	17	12	14	19	11	18	13	17
ELEVATION	9	16	4	5	18	17	4	5	1	13	16	9
FRE3	11	12	3	2	11	9	7	15	4	4	5	7
ORDER	19	19	19	18	15	18	19	17	10	19	17	15
SEGFLOWSTA	3	4	9	6	10	13	2	2	5	3	8	2
SEGJANAIRT	16	5	14	14	13	8	15	18	2	8	9	14
SEGMINTNOR	14	15	16	10	12	7	9	12	12	9	15	11
SEGRIPSHAD	15	6	12	11	7	16	16	13	15	11	18	19
SpecMALF	4	8	6	8	4	6	10	9	17	12	11	8
SpecMeanF	6	9	2	4	3	2	3	16	18	16	4	4
USAVGSLOPE	1	2	1	3	1	1	5	4	9	7	3	13
USCALCIUM	7	11	10	9	14	10	6	11	16	17	14	6
USHARDNESS	10	10	7	1	9	14	12	6	6	14	12	16
USPHOSPHOR	5	14	17	15	19	11	13	3	7	15	7	5

Table A5.4. Ranks of relative importance of variables from random forest models of macroinvertebrate traits in response to reach-scale stressors and environmental descriptors. The R^2 value for each trait model is also shown.

trait	1a	2c	3b	3c	5a	6b	6c	7b	9a	10b	11b	16a	16b
R^2	0.57	0.57	0.62	0.66	0.56	0.50	0.61	0.64	0.56	0.34	0.63	0.60	0.64
CHLA	11	14	3	5	2	1	2	10	2	4	1	15	3
DIN	14	10	9	14	15	5	6	9	7	6	8	6	12
DRP	1	1	4	2	12	17	12	13	18	3	5	8	9
instreamVis	5	5	10	8	3	18	14	3	5	1	14	7	13
maxrateToQ50	12	18	6	11	7	13	15	12	14	12	12	11	16
DSDIST2COA	6	11	15	13	17	16	16	18	19	18	13	16	14
ELEVATION	17	19	17	16	19	7	7	11	10	17	16	12	10
FRE3	7	13	12	10	8	8	11	2	9	13	10	14	11
ORDER_	19	3	19	17	18	19	17	15	17	19	18	19	17
SEGFLOWSTA	4	7	5	4	5	3	4	8	8	10	6	3	2
SEGJANAIRT	18	15	11	9	9	12	18	19	11	14	15	18	18
SEGMINTNOR	13	9	13	12	13	9	13	14	15	11	9	13	8
SEGRIPSHAD	10	16	18	19	4	15	9	17	12	16	19	9	19
SpecMALF	3	6	2	1	11	6	3	6	4	7	4	4	6
SpecMeanF	8	2	8	7	16	4	5	4	3	8	11	5	5
USAVGSLOPE	2	8	1	3	1	2	1	1	1	2	2	1	1
USCALCIUM	16	12	14	15	14	10	10	16	16	9	7	10	7
USHARDNESS	15	17	16	18	10	11	8	5	13	15	17	2	15
USPHOSPHOR	9	4	7	6	6	14	19	7	6	5	3	17	4

Table A5.5. Multiple regression coefficient of determination (R^2), predictor variable coefficients, standard error (S.E.) and 95% confidence intervals (CI) for each metric and trait catchment-scale model. *excluding Hydoptilidae.

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
AMDI	0.33	T1NativeVeg	0.30	0.01	0.00000	0.27	0.33
		T1ExoticVeg	0.13	0.01	0.00000	0.11	0.15
		T1Urban	-0.15	0.01	0.00000	-0.17	-0.13
		ORDER_	-0.13	0.01	0.00000	-0.16	-0.11
		ELEVATION	0.29	0.01	0.00000	0.27	0.31
		SEGRIPSHAD	0.04	0.01	0.00240	0.01	0.07
		SEGMINTNOR	0.10	0.01	0.00000	0.08	0.12
		USAVGSLOPE	-0.10	0.02	0.00000	-0.13	-0.07
		USPHOSPHOR	0.10	0.01	0.00000	0.08	0.12
		USHARDNESS	0.12	0.01	0.00000	0.09	0.14
		SpecMALF	0.09	0.01	0.00000	0.07	0.12
EPTrich*	0.37	T1NativeVeg	0.37	0.01	0.00000	0.35	0.40
		T1ExoticVeg	0.05	0.01	0.00000	0.03	0.07
		T1Urban	-0.13	0.01	0.00000	-0.15	-0.11
		ELEVATION	0.23	0.01	0.00000	0.20	0.25
		DSDIST2COA	-0.05	0.01	0.00011	-0.08	-0.03
		SEGJANAIRT	-0.13	0.01	0.00000	-0.15	-0.11
		SEGMINTNOR	0.07	0.01	0.00000	0.05	0.10
		USAVGSLOPE	-0.10	0.02	0.00000	-0.13	-0.07
		USPHOSPHOR	0.07	0.01	0.00000	0.05	0.09
		USHARDNESS	0.12	0.01	0.00000	0.10	0.14
		SEGFLOWSTA	0.04	0.01	0.00140	0.02	0.06
		SpecMALF	0.13	0.01	0.00000	0.11	0.16
		FRE3	-0.04	0.01	0.00021	-0.07	-0.02
LIFENZ	0.58	T1ExoticVeg	0.07	0.02	0.00000	0.04	0.11
		T1Urban	-0.06	0.01	0.00007	-0.09	-0.03
		maxrateToQ50	-0.06	0.02	0.00012	-0.09	-0.03
		ELEVATION	0.28	0.02	0.00000	0.23	0.33
		DSDIST2COA	-0.24	0.03	0.00000	-0.30	-0.19
		SEGRIPSHAD	-0.10	0.02	0.00000	-0.14	-0.06
		SEGJANAIRT	-0.27	0.02	0.00000	-0.31	-0.23
		USPHOSPHOR	0.17	0.02	0.00000	0.14	0.21
		USHARDNESS	0.22	0.02	0.00000	0.19	0.25
		SpecMALF	0.33	0.02	0.00000	0.29	0.37
MCI_hb	0.50	T1NativeVeg	0.27	0.01	0.00000	0.24	0.29
		T1ExoticVeg	0.08	0.01	0.00000	0.06	0.09
		T2PastoralHeavy	-0.07	0.01	0.00000	-0.09	-0.05
		T1Urban	-0.11	0.01	0.00000	-0.12	-0.09
		maxrateToQ50	0.03	0.01	0.00007	0.02	0.05

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
		ORDER_	-0.04	0.01	0.00001	-0.06	-0.02
		ELEVATION	0.22	0.01	0.00000	0.19	0.24
		DSDIST2COA	-0.07	0.01	0.00000	-0.10	-0.05
		SEGJANAIRT	-0.15	0.01	0.00000	-0.17	-0.13
		USAVGSLOPE	0.07	0.01	0.00000	0.04	0.10
		USCALCIUM	-0.03	0.01	0.00099	-0.04	-0.01
		USHARDNESS	0.10	0.01	0.00000	0.08	0.12
		SEGFLOWSTA	0.04	0.01	0.00117	0.01	0.06
		SpecMALF	0.17	0.01	0.00000	0.15	0.20
		FRE3	0.06	0.01	0.00000	0.04	0.08
pEPTabund*	0.38	T1NativeVeg	0.23	0.01	0.00000	0.21	0.26
		T1Urban	-0.10	0.01	0.00000	-0.12	-0.09
		maxrateToQ50	0.03	0.01	0.00201	0.01	0.05
		ORDER_	0.10	0.01	0.00000	0.08	0.12
		ELEVATION	0.25	0.01	0.00000	0.22	0.28
		DSDIST2COA	-0.13	0.01	0.00000	-0.15	-0.10
		SEGJANAIRT	-0.16	0.01	0.00000	-0.18	-0.14
		USAVGSLOPE	0.11	0.01	0.00000	0.08	0.13
		USCALCIUM	-0.06	0.01	0.00000	-0.08	-0.04
		USPHOSPHOR	0.07	0.01	0.00000	0.05	0.09
		USHARDNESS	0.07	0.01	0.00000	0.04	0.09
		SpecMALF	0.11	0.01	0.00000	0.09	0.13
		FRE3	-0.04	0.01	0.00065	-0.06	-0.02
pEPTrich*	0.41	T1NativeVeg	0.35	0.01	0.00000	0.32	0.37
		T1ExoticVeg	0.06	0.01	0.00000	0.04	0.08
		T2PastoralHeavy	0.06	0.01	0.00000	0.04	0.09
		T1Urban	-0.10	0.01	0.00000	-0.11	-0.08
		maxrateToQ50	0.03	0.01	0.00016	0.02	0.05
		ORDER_	0.09	0.01	0.00000	0.07	0.11
		ELEVATION	0.16	0.01	0.00000	0.14	0.19
		DSDIST2COA	-0.05	0.01	0.00015	-0.07	-0.02
		SEGJANAIRT	-0.22	0.01	0.00000	-0.24	-0.20
		USPHOSPHOR	0.04	0.01	0.00007	0.02	0.06
		USHARDNESS	0.14	0.01	0.00000	0.12	0.16
		SpecMALF	0.20	0.01	0.00000	0.18	0.22
QMCI_hb	0.27	T1NativeVeg	0.34	0.01	0.00000	0.31	0.37
		T1Urban	-0.06	0.01	0.00000	-0.09	-0.04
		maxrateToQ50	0.05	0.01	0.00005	0.02	0.07
		ELEVATION	0.09	0.01	0.00000	0.06	0.11
		SEGJANAIRT	-0.20	0.01	0.00000	-0.23	-0.18
		USAVGSLOPE	0.16	0.02	0.00000	0.13	0.19

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
		USPHOSPHOR	-0.11	0.01	0.00000	-0.14	-0.09
		FRE3	-0.10	0.02	0.00000	-0.13	-0.07
QUCI	0.40	T1NativeVeg	0.35	0.02	0.00000	0.32	0.38
		T1ExoticVeg	0.08	0.01	0.00000	0.06	0.10
		T1Urban	-0.09	0.01	0.00000	-0.11	-0.07
		maxrateToQ50	0.05	0.01	0.00001	0.03	0.07
		ORDER_	0.06	0.02	0.00005	0.03	0.09
		ELEVATION	0.23	0.02	0.00000	0.20	0.26
		DSDIST2COA	-0.10	0.02	0.00000	-0.13	-0.07
		SEGRIPSHAD	0.06	0.02	0.00005	0.03	0.09
		SEGJANAIRT	-0.18	0.01	0.00000	-0.20	-0.15
		USAVGSLOPE	0.15	0.02	0.00000	0.12	0.19
		USPHOSPHOR	-0.07	0.01	0.00000	-0.09	-0.05
		SpecMALF	0.08	0.01	0.00000	0.05	0.10
		FRE3	-0.14	0.01	0.00000	-0.17	-0.11
UCI	0.51	T1NativeVeg	0.33	0.01	0.00000	0.31	0.35
		T1ExoticVeg	0.13	0.01	0.00000	0.12	0.15
		T1Urban	-0.08	0.01	0.00000	-0.09	-0.06
		maxrateToQ50	0.03	0.01	0.00026	0.01	0.04
		ORDER_	0.11	0.01	0.00000	0.09	0.12
		ELEVATION	0.25	0.01	0.00000	0.23	0.28
		DSDIST2COA	-0.09	0.01	0.00000	-0.12	-0.07
		SEGJANAIRT	-0.05	0.01	0.00000	-0.07	-0.03
		USPHOSPHOR	0.04	0.01	0.00000	0.03	0.06
		USHARDNESS	0.13	0.01	0.00000	0.11	0.15
		SEGFLOWSTA	0.05	0.01	0.00000	0.03	0.07
		SpecMALF	0.25	0.01	0.00000	0.23	0.27
		FRE3	0.07	0.01	0.00000	0.05	0.09
Simpsons	0.12	T1NativeVeg	0.22	0.02	0.00000	0.20	0.25
		T1ExoticVeg	0.13	0.01	0.00000	0.10	0.15
		T1Urban	-0.11	0.01	0.00000	-0.13	-0.09
		ORDER_	-0.15	0.02	0.00000	-0.19	-0.12
		ELEVATION	0.10	0.02	0.00000	0.06	0.13
		DSDIST2COA	-0.07	0.02	0.00000	-0.11	-0.04
		SEGRIPSHAD	0.09	0.02	0.00000	0.06	0.12
		SEGJANAIRT	-0.04	0.01	0.00067	-0.07	-0.02
		USAVGSLOPE	-0.11	0.01	0.00000	-0.14	-0.08
		USCALCIUM	-0.05	0.01	0.00007	-0.07	-0.02
		USPHOSPHOR	0.12	0.01	0.00000	0.10	0.15
		FRE3	-0.08	0.01	0.00000	-0.10	-0.05

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
totRich	0.16	T1NativeVeg	0.13	0.02	0.00000	0.09	0.16
		T1ExoticVeg	0.05	0.01	0.00001	0.03	0.07
		T2PastoralHeavy	-0.06	0.02	0.00046	-0.09	-0.03
		T1Urban	-0.12	0.01	0.00000	-0.15	-0.10
		ORDER_	-0.14	0.02	0.00000	-0.17	-0.11
		ELEVATION	0.19	0.01	0.00000	0.17	0.22
		SEGRIPSHAD	0.05	0.02	0.00188	0.02	0.08
		SEGJANAIRT	0.08	0.01	0.00000	0.06	0.11
		SEGMINTNOR	0.12	0.01	0.00000	0.10	0.14
		USAVGSLOPE	-0.12	0.02	0.00000	-0.15	-0.08
		USPHOSPHOR	0.11	0.01	0.00000	0.09	0.13
		USHARDNESS	0.05	0.01	0.00010	0.03	0.08
		SEGFLOWSTA	0.05	0.01	0.00002	0.03	0.08
1b	0.24	T1NativeVeg	0.08	0.02	0.00000	0.04	0.11
		T1ExoticVeg	0.13	0.01	0.00000	0.11	0.15
		T2PastoralHeavy	0.06	0.02	0.00007	0.03	0.10
		T1Urban	-0.13	0.01	0.00000	-0.15	-0.11
		ELEVATION	0.10	0.01	0.00000	0.08	0.12
		SEGRIPSHAD	-0.14	0.01	0.00000	-0.16	-0.11
		SEGJANAIRT	-0.14	0.01	0.00000	-0.17	-0.12
		SEGMINTNOR	-0.07	0.01	0.00000	-0.09	-0.05
		USAVGSLOPE	0.13	0.02	0.00000	0.10	0.17
		USCALCIUM	-0.10	0.01	0.00000	-0.12	-0.08
		USPHOSPHOR	0.12	0.01	0.00000	0.10	0.14
		USHARDNESS	0.06	0.01	0.00003	0.03	0.08
		SpecMALF	0.05	0.01	0.00003	0.03	0.08
1e	0.22	T1NativeVeg	0.24	0.02	0.00000	0.21	0.27
		T1ExoticVeg	0.11	0.01	0.00000	0.09	0.13
		T2PastoralHeavy	0.07	0.02	0.00000	0.04	0.11
		ELEVATION	0.25	0.02	0.00000	0.22	0.29
		DSDIST2COA	-0.12	0.02	0.00000	-0.15	-0.09
		SEGRIPSHAD	0.08	0.01	0.00000	0.05	0.10
		SEGJANAIRT	0.08	0.01	0.00000	0.06	0.11
		USAVGSLOPE	0.06	0.02	0.00010	0.03	0.10
		USPHOSPHOR	-0.05	0.01	0.00001	-0.08	-0.03
		USHARDNESS	0.05	0.01	0.00027	0.02	0.07
		SEGFLOWSTA	-0.05	0.01	0.00003	-0.08	-0.03
		SpecMALF	0.11	0.01	0.00000	0.08	0.14
		FRE3	0.08	0.01	0.00000	0.06	0.10
2a	0.46	T1NativeVeg	-0.24	0.01	0.00000	-0.27	-0.21
		T1ExoticVeg	-0.06	0.01	0.00000	-0.08	-0.04

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
		T2PastoralHeavy	0.06	0.01	0.00001	0.03	0.09
		T1Urban	0.07	0.01	0.00000	0.06	0.09
		ORDER_	-0.04	0.01	0.00101	-0.06	-0.02
		ELEVATION	-0.26	0.01	0.00000	-0.28	-0.23
		DSDIST2COA	0.12	0.01	0.00000	0.10	0.15
		SEGRIPSHAD	-0.05	0.01	0.00005	-0.08	-0.03
		SEGJANAIRT	0.04	0.01	0.00037	0.02	0.06
		SEGMINTNOR	0.04	0.01	0.00002	0.02	0.06
		USAVGSLOPE	-0.07	0.01	0.00001	-0.09	-0.04
		USCALCIUM	0.07	0.01	0.00000	0.05	0.09
		USPHOSPHOR	-0.12	0.01	0.00000	-0.14	-0.10
		USHARDNESS	-0.09	0.01	0.00000	-0.11	-0.06
		SEGFLOWSTA	-0.07	0.01	0.00000	-0.10	-0.05
		SpecMALF	-0.17	0.01	0.00000	-0.20	-0.15
2b	0.39	T1NativeVeg	0.17	0.01	0.00000	0.14	0.20
		T1ExoticVeg	0.06	0.01	0.00000	0.04	0.08
		T2PastoralHeavy	-0.04	0.01	0.00247	-0.07	-0.01
		T1Urban	-0.07	0.01	0.00000	-0.09	-0.05
		ORDER_	0.05	0.01	0.00001	0.03	0.07
		ELEVATION	0.21	0.01	0.00000	0.18	0.23
		DSDIST2COA	-0.10	0.01	0.00000	-0.13	-0.08
		SEGJANAIRT	-0.10	0.01	0.00000	-0.12	-0.08
		SEGMINTNOR	-0.11	0.01	0.00000	-0.13	-0.09
		USAVGSLOPE	0.09	0.02	0.00000	0.06	0.12
		USCALCIUM	-0.10	0.01	0.00000	-0.12	-0.08
		USPHOSPHOR	0.18	0.01	0.00000	0.15	0.20
		USHARDNESS	0.07	0.01	0.00000	0.05	0.10
		SEGFLOWSTA	0.07	0.01	0.00000	0.05	0.09
		SpecMALF	0.15	0.01	0.00000	0.12	0.17
2d	0.29	T1NativeVeg	0.31	0.02	0.00000	0.28	0.34
		T1ExoticVeg	0.05	0.01	0.00000	0.03	0.07
		T2PastoralHeavy	-0.07	0.01	0.00001	-0.09	-0.04
		T1Urban	-0.07	0.01	0.00000	-0.08	-0.05
		ELEVATION	0.19	0.02	0.00000	0.16	0.22
		DSDIST2COA	-0.16	0.01	0.00000	-0.18	-0.13
		SEGRIPSHAD	0.13	0.01	0.00000	0.11	0.15
		SEGJANAIRT	0.12	0.01	0.00000	0.10	0.14
		SEGMINTNOR	0.08	0.01	0.00000	0.06	0.10
		USAVGSLOPE	-0.12	0.01	0.00000	-0.15	-0.09
		SEGFLOWSTA	0.11	0.01	0.00000	0.08	0.13
		SpecMALF	0.09	0.01	0.00000	0.06	0.12

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
3b	0.37	T1NativeVeg	0.24	0.01	0.00000	0.21	0.26
		T1ExoticVeg	0.03	0.01	0.00039	0.01	0.05
		T2PastoralHeavy	-0.05	0.01	0.00024	-0.07	-0.02
		T1Urban	-0.11	0.01	0.00000	-0.13	-0.09
		ORDER_	0.07	0.01	0.00000	0.05	0.09
		ELEVATION	0.19	0.01	0.00000	0.16	0.22
		DSDIST2COA	-0.10	0.01	0.00000	-0.13	-0.07
		SEGJANAIRT	-0.18	0.01	0.00000	-0.20	-0.16
		USCALCIUM	-0.06	0.01	0.00000	-0.08	-0.04
		USPHOSPHOR	0.05	0.01	0.00001	0.03	0.07
		USHARDNESS	0.07	0.01	0.00000	0.05	0.10
		SpecMALF	0.18	0.01	0.00000	0.16	0.21
3c	0.39	T1NativeVeg	-0.24	0.01	0.00000	-0.26	-0.21
		T1ExoticVeg	-0.04	0.01	0.00000	-0.06	-0.02
		T2PastoralHeavy	0.05	0.01	0.00002	0.03	0.08
		T1Urban	0.10	0.01	0.00000	0.08	0.12
		maxrateToQ50	-0.03	0.01	0.00094	-0.05	-0.01
		ORDER_	-0.07	0.01	0.00000	-0.09	-0.05
		ELEVATION	-0.23	0.01	0.00000	-0.26	-0.20
		DSDIST2COA	0.13	0.01	0.00000	0.11	0.16
		SEGJANAIRT	0.17	0.01	0.00000	0.15	0.19
		USCALCIUM	0.04	0.01	0.00001	0.02	0.06
		USHARDNESS	-0.09	0.01	0.00000	-0.11	-0.07
		SpecMALF	-0.19	0.01	0.00000	-0.21	-0.17
		FRE3	-0.05	0.01	0.00000	-0.07	-0.03
4a	0.53	T1NativeVeg	0.20	0.01	0.00000	0.18	0.23
		T1ExoticVeg	0.10	0.01	0.00000	0.08	0.12
		T2PastoralHeavy	-0.06	0.01	0.00000	-0.08	-0.04
		T1Urban	-0.06	0.01	0.00000	-0.08	-0.05
		maxrateToQ50	0.03	0.01	0.00004	0.02	0.05
		ORDER_	0.08	0.01	0.00000	0.06	0.10
		ELEVATION	0.29	0.01	0.00000	0.26	0.31
		DSDIST2COA	-0.12	0.01	0.00000	-0.15	-0.10
		SEGJANAIRT	-0.10	0.01	0.00000	-0.11	-0.08
		SEGMINTNOR	-0.03	0.01	0.00216	-0.05	-0.01
		USAVGSLOPE	0.18	0.01	0.00000	0.15	0.21
		USPHOSPHOR	0.03	0.01	0.00245	0.01	0.04
		USHARDNESS	0.04	0.01	0.00018	0.02	0.06
		SpecMALF	0.24	0.01	0.00000	0.23	0.26
		FRE3	0.03	0.01	0.00132	0.01	0.05

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
4c	0.53	T1NativeVeg	-0.19	0.01	0.00000	-0.22	-0.17
		T1ExoticVeg	-0.09	0.01	0.00000	-0.10	-0.07
		T2PastoralHeavy	0.07	0.01	0.00000	0.05	0.09
		T1Urban	0.06	0.01	0.00000	0.04	0.08
		maxrateToQ50	-0.03	0.01	0.00002	-0.05	-0.02
		ORDER_	-0.08	0.01	0.00000	-0.10	-0.07
		ELEVATION	-0.29	0.01	0.00000	-0.31	-0.26
		DSDIST2COA	0.13	0.01	0.00000	0.11	0.16
		SEGJANAIRT	0.10	0.01	0.00000	0.09	0.12
		SEGMINTNOR	0.04	0.01	0.00050	0.02	0.06
		USAVGSLOPE	-0.17	0.01	0.00000	-0.20	-0.14
		USHARDNESS	-0.04	0.01	0.00016	-0.06	-0.02
		SpecMALF	-0.25	0.01	0.00000	-0.27	-0.23
		FRE3	-0.04	0.01	0.00036	-0.06	-0.02
5a	0.40	T1ExoticVeg	-0.05	0.01	0.00000	-0.07	-0.04
		T1Urban	-0.05	0.01	0.00000	-0.06	-0.03
		maxrateToQ50	0.03	0.01	0.00036	0.01	0.05
		ORDER_	0.16	0.01	0.00000	0.14	0.19
		DSDIST2COA	-0.09	0.01	0.00000	-0.11	-0.07
		SEGRIPSHAD	-0.05	0.01	0.00001	-0.08	-0.03
		SEGJANAIRT	-0.36	0.01	0.00000	-0.38	-0.34
		USAVGSLOPE	0.31	0.01	0.00000	0.29	0.34
		USCALCIUM	-0.03	0.01	0.00029	-0.05	-0.01
		USHARDNESS	0.08	0.01	0.00000	0.05	0.10
		SpecMALF	0.10	0.01	0.00000	0.08	0.12
5d	0.38	T1NativeVeg	-0.20	0.01	0.00000	-0.22	-0.17
		T1ExoticVeg	-0.13	0.01	0.00000	-0.15	-0.11
		T1Urban	0.06	0.01	0.00000	0.04	0.07
		ORDER_	-0.06	0.01	0.00000	-0.08	-0.04
		ELEVATION	-0.24	0.01	0.00000	-0.26	-0.21
		DSDIST2COA	0.17	0.01	0.00000	0.15	0.20
		SEGMINTNOR	0.10	0.01	0.00000	0.08	0.12
		USAVGSLOPE	-0.15	0.02	0.00000	-0.18	-0.12
		USCALCIUM	0.05	0.01	0.00000	0.03	0.07
		USPHOSPHOR	-0.14	0.01	0.00000	-0.16	-0.12
		USHARDNESS	-0.05	0.01	0.00000	-0.08	-0.03
		SEGFLOWSTA	-0.07	0.01	0.00000	-0.09	-0.05
		SpecMALF	-0.15	0.01	0.00000	-0.18	-0.13

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
6c	0.46	T1NativeVeg	0.22	0.01	0.00000	0.19	0.24
		T1ExoticVeg	0.10	0.01	0.00000	0.09	0.12
		T1Urban	-0.06	0.01	0.00000	-0.07	-0.04
		maxrateToQ50	0.03	0.01	0.00008	0.02	0.05
		ORDER_	0.04	0.01	0.00005	0.02	0.06
		ELEVATION	0.26	0.01	0.00000	0.23	0.28
		DSDIST2COA	-0.16	0.01	0.00000	-0.19	-0.14
		SEGJANAIRT	-0.14	0.01	0.00000	-0.16	-0.13
		SEGMINTNOR	-0.05	0.01	0.00001	-0.07	-0.03
		USAVGSLOPE	0.20	0.01	0.00000	0.17	0.23
		USHARDNESS	0.05	0.01	0.00000	0.03	0.07
		SpecMALF	0.19	0.01	0.00000	0.17	0.21
		FRE3	0.05	0.01	0.00000	0.03	0.07
7a	0.41	T1NativeVeg	0.23	0.01	0.00000	0.21	0.26
		T1ExoticVeg	0.07	0.01	0.00000	0.05	0.09
		T2PastoralHeavy	-0.06	0.01	0.00000	-0.09	-0.04
		T1Urban	-0.12	0.01	0.00000	-0.13	-0.10
		ELEVATION	0.22	0.01	0.00000	0.20	0.25
		DSDIST2COA	-0.07	0.01	0.00000	-0.09	-0.04
		SEGJANAIRT	-0.19	0.01	0.00000	-0.21	-0.17
		USCALCIUM	-0.05	0.01	0.00000	-0.07	-0.03
		USPHOSPHOR	0.11	0.01	0.00000	0.09	0.13
		USHARDNESS	0.10	0.01	0.00000	0.08	0.12
		SEGFLOWSTA	0.05	0.01	0.00003	0.03	0.07
		SpecMALF	0.13	0.01	0.00000	0.11	0.16
7b	0.45	T1NativeVeg	-0.22	0.01	0.00000	-0.24	-0.19
		T1ExoticVeg	-0.08	0.01	0.00000	-0.10	-0.07
		T2PastoralHeavy	0.07	0.01	0.00000	0.05	0.09
		T1Urban	0.10	0.01	0.00000	0.08	0.11
		ELEVATION	-0.26	0.01	0.00000	-0.29	-0.24
		DSDIST2COA	0.08	0.01	0.00000	0.06	0.11
		SEGJANAIRT	0.16	0.01	0.00000	0.14	0.18
		USCALCIUM	0.04	0.01	0.00001	0.02	0.06
		USPHOSPHOR	-0.06	0.01	0.00000	-0.08	-0.04
		USHARDNESS	-0.09	0.01	0.00000	-0.11	-0.07
		SEGFLOWSTA	-0.04	0.01	0.00133	-0.06	-0.01
		SpecMALF	-0.20	0.01	0.00000	-0.22	-0.17
		FRE3	-0.04	0.01	0.00003	-0.06	-0.02

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
8c	0.46	T1NativeVeg	-0.10	0.01	0.00000	-0.13	-0.07
		T1ExoticVeg	-0.04	0.01	0.00000	-0.06	-0.03
		T2PastoralHeavy	0.14	0.01	0.00000	0.12	0.17
		T1Urban	0.05	0.01	0.00000	0.03	0.06
		ORDER_	-0.07	0.01	0.00000	-0.09	-0.05
		ELEVATION	-0.24	0.01	0.00000	-0.27	-0.22
		DSDIST2COA	0.10	0.01	0.00000	0.07	0.12
		SEGJANAIRT	0.11	0.01	0.00000	0.09	0.13
		SEGMINTNOR	0.03	0.01	0.00217	0.01	0.05
		USAVGSLOPE	-0.19	0.01	0.00000	-0.21	-0.16
		USCALCIUM	0.05	0.01	0.00000	0.03	0.07
		USPHOSPHOR	-0.06	0.01	0.00000	-0.08	-0.04
		SpecMALF	-0.23	0.01	0.00000	-0.25	-0.21
11b	0.34	T1NativeVeg	0.25	0.01	0.00000	0.23	0.28
		T1Urban	-0.04	0.01	0.00000	-0.06	-0.03
		maxrateToQ50	0.05	0.01	0.00000	0.03	0.06
		ORDER_	0.05	0.01	0.00035	0.02	0.07
		ELEVATION	0.18	0.01	0.00000	0.15	0.21
		DSDIST2COA	-0.05	0.01	0.00130	-0.08	-0.02
		SEGRIPSHAD	0.05	0.01	0.00008	0.03	0.08
		SEGJANAIRT	-0.10	0.01	0.00000	-0.12	-0.08
		SEGMINTNOR	0.11	0.01	0.00000	0.09	0.13
		USAVGSLOPE	0.15	0.02	0.00000	0.12	0.18
		USPHOSPHOR	-0.07	0.01	0.00000	-0.09	-0.05
		USHARDNESS	0.04	0.01	0.00154	0.01	0.06
		SpecMALF	0.14	0.01	0.00000	0.12	0.16
12a	0.26	T1NativeVeg	-0.13	0.01	0.00000	-0.16	-0.10
		T1ExoticVeg	-0.08	0.01	0.00000	-0.10	-0.06
		maxrateToQ50	0.03	0.01	0.00211	0.01	0.05
		ORDER_	0.13	0.01	0.00000	0.11	0.16
		ELEVATION	-0.16	0.02	0.00000	-0.19	-0.13
		DSDIST2COA	-0.08	0.02	0.00000	-0.11	-0.05
		SEGJANAIRT	-0.45	0.01	0.00000	-0.47	-0.43
		SEGMINTNOR	-0.04	0.01	0.00210	-0.06	-0.01
		USAVGSLOPE	0.10	0.01	0.00000	0.07	0.13
		USPHOSPHOR	-0.04	0.01	0.00007	-0.07	-0.02
		SEGFLOWSTA	-0.10	0.01	0.00000	-0.12	-0.07
		SpecMALF	0.07	0.01	0.00000	0.04	0.09

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
13c	0.35	T1NativeVeg	-0.12	0.01	0.00000	-0.15	-0.10
		T1ExoticVeg	-0.08	0.01	0.00000	-0.10	-0.06
		T1Urban	0.06	0.01	0.00000	0.04	0.08
		ORDER_	-0.18	0.01	0.00000	-0.20	-0.15
		ELEVATION	-0.34	0.01	0.00000	-0.37	-0.31
		DSDIST2COA	0.08	0.01	0.00000	0.05	0.11
		SEGRIPSHAD	0.13	0.01	0.00000	0.10	0.15
		SEGJANAIRT	-0.19	0.01	0.00000	-0.21	-0.17
		USAVGSLOPE	-0.13	0.01	0.00000	-0.15	-0.10
		USCALCIUM	0.05	0.01	0.00000	0.03	0.07
		USPHOSPHOR	-0.03	0.01	0.00179	-0.06	-0.01
		SpecMALF	-0.18	0.01	0.00000	-0.20	-0.15
		FRE3	-0.09	0.01	0.00000	-0.11	-0.07
14c	0.08	T2PastoralHeavy	0.07	0.01	0.00000	0.05	0.10
		ORDER_	0.19	0.01	0.00000	0.17	0.21
		DSDIST2COA	0.06	0.01	0.00000	0.03	0.08
		SEGJANAIRT	-0.06	0.01	0.00000	-0.08	-0.03
		SEGMINTNOR	0.14	0.01	0.00000	0.12	0.17
		USAVGSLOPE	0.13	0.01	0.00000	0.10	0.16
		USCALCIUM	0.13	0.01	0.00000	0.11	0.15
		USPHOSPHOR	-0.05	0.01	0.00030	-0.07	-0.02
15b	0.08	T1NativeVeg	0.24	0.01	0.00000	0.21	0.27
		T1ExoticVeg	0.05	0.01	0.00002	0.02	0.07
		T2PastoralHeavy	0.13	0.02	0.00000	0.10	0.16
		T1Urban	-0.05	0.01	0.00003	-0.07	-0.02
		ORDER_	0.09	0.01	0.00000	0.07	0.11
		ELEVATION	0.06	0.01	0.00000	0.04	0.09
		SEGMINTNOR	0.16	0.01	0.00000	0.14	0.19
		USHARDNESS	0.05	0.01	0.00010	0.02	0.07
		FRE3	-0.06	0.01	0.00002	-0.09	-0.03
15c	0.10	T1NativeVeg	-0.12	0.01	0.00000	-0.15	-0.10
		ORDER_	0.21	0.02	0.00000	0.17	0.24
		ELEVATION	0.12	0.01	0.00000	0.10	0.15
		SEGRIPSHAD	-0.10	0.02	0.00000	-0.13	-0.07
		SEGJANAIRT	0.08	0.01	0.00000	0.05	0.10
		SEGMINTNOR	0.10	0.01	0.00000	0.08	0.13
		USPHOSPHOR	-0.11	0.01	0.00000	-0.14	-0.09
		SEGFLOWSTA	-0.10	0.01	0.00000	-0.12	-0.07
		FRE3	0.16	0.01	0.00000	0.13	0.19

Table A5.5, continued

Response	R^2	Term	Coefficient	S.E.	*P* - value	2.5% CI	97.5% CI
16a	0.48	T1NativeVeg	-0.22	0.01	0.00000	-0.24	-0.19
		T1ExoticVeg	-0.09	0.01	0.00000	-0.10	-0.07
		T2PastoralHeavy	0.07	0.01	0.00000	0.04	0.10
		T1Urban	0.05	0.01	0.00000	0.04	0.07
		ORDER_	-0.04	0.01	0.00062	-0.07	-0.02
		ELEVATION	-0.25	0.01	0.00000	-0.27	-0.22
		DSDIST2COA	0.10	0.01	0.00000	0.08	0.13
		SEGRIPSHAD	-0.06	0.01	0.00001	-0.09	-0.03
		SEGJANAIRT	0.06	0.01	0.00000	0.04	0.08
		SEGMINTNOR	0.05	0.01	0.00000	0.03	0.07
		USAVGSLOPE	-0.15	0.01	0.00000	-0.18	-0.12
		USCALCIUM	0.04	0.01	0.00005	0.02	0.06
		USPHOSPHOR	-0.11	0.01	0.00000	-0.13	-0.09
		USHARDNESS	-0.04	0.01	0.00125	-0.06	-0.01
		SEGFLOWSTA	-0.06	0.01	0.00000	-0.08	-0.04
		SpecMALF	-0.19	0.01	0.00000	-0.21	-0.16
16b	0.48	T1NativeVeg	0.20	0.01	0.00000	0.18	0.23
		T1ExoticVeg	0.04	0.01	0.00000	0.03	0.06
		T2PastoralHeavy	-0.09	0.01	0.00000	-0.12	-0.07
		T1Urban	-0.08	0.01	0.00000	-0.10	-0.06
		maxrateToQ50	0.03	0.01	0.00137	0.01	0.04
		ELEVATION	0.23	0.01	0.00000	0.21	0.25
		SEGRIPSHAD	0.08	0.01	0.00000	0.06	0.10
		SEGJANAIRT	-0.03	0.01	0.00015	-0.05	-0.02
		SEGMINTNOR	0.12	0.01	0.00000	0.10	0.14
		USAVGSLOPE	0.15	0.01	0.00000	0.13	0.18
		SEGFLOWSTA	0.04	0.01	0.00045	0.02	0.06
		SpecMALF	0.21	0.01	0.00000	0.18	0.23

Table A5.6. Multiple regression coefficient of determination (R^2), predictor variable coefficients, standard error (S.E.) and 95% confidence intervals (CI) for each metric and trait reach-scale model. *excluding Hydoptilidae.

Response variable	Type	R^2	Predictors	Estimate	S.E.	*P* - value	2.5% CI	97.5% CI
MCI_hb	metric	0.452	instreamVis	-0.09	0.04	0.011	-0.17	-0.02
			CHLA	-0.20	0.04	0.000	-0.27	-0.13
			maxrateToQ50	-0.10	0.03	0.003	-0.17	-0.04
			ORDER_	-0.14	0.04	0.000	-0.21	-0.07
			DSDIST2COA	0.15	0.05	0.002	0.06	0.25
			SEGMINTNOR	0.24	0.05	0.000	0.15	0.33
			USAVGSLOPE	0.34	0.04	0.000	0.26	0.42
			SpecMALF	0.27	0.04	0.000	0.19	0.34
EPTrich	metric	0.226	maxrateToQ50	-0.14	0.04	0.000	-0.22	-0.06
			DSDIST2COA	0.15	0.06	0.006	0.04	0.26
			SEGRIPSHAD	-0.16	0.04	0.000	-0.24	-0.07
			SEGJANAIRT	-0.37	0.05	0.000	-0.48	-0.26
			SEGMINTNOR	0.20	0.06	0.000	0.09	0.32
			USPHOSPHOR	-0.27	0.06	0.000	-0.38	-0.15
			SpecMALF	0.22	0.05	0.000	0.13	0.31
pEPTrich	metric	0.349	instreamVis	-0.15	0.04	0.000	-0.22	-0.07
			CHLA	-0.14	0.04	0.000	-0.21	-0.06
			maxrateToQ50	-0.11	0.04	0.003	-0.18	-0.04
			SEGRIPSHAD	-0.35	0.04	0.000	-0.43	-0.28
			SEGJANAIRT	-0.27	0.05	0.000	-0.37	-0.18
			USPHOSPHOR	-0.22	0.05	0.000	-0.32	-0.12
			SpecMALF	0.23	0.04	0.000	0.15	0.31
pEPTabund	metric	0.335	instreamVis	-0.15	0.04	0.000	-0.23	-0.07
			CHLA	-0.17	0.04	0.000	-0.25	-0.09
			DSDIST2COA	0.14	0.05	0.006	0.04	0.24
			SEGRIPSHAD	-0.24	0.04	0.000	-0.32	-0.16
			SEGJANAIRT	-0.16	0.04	0.000	-0.24	-0.08
			SEGMINTNOR	0.21	0.05	0.000	0.10	0.31
			USAVGSLOPE	0.27	0.05	0.000	0.18	0.36
			USCALCIUM	-0.13	0.04	0.002	-0.21	-0.05
sed_MCI_like	metric	0.416	instreamVis	-0.21	0.04	0.000	-0.29	-0.14
			maxrateToQ50	-0.11	0.04	0.002	-0.18	-0.04
			ORDER_	-0.20	0.05	0.000	-0.30	-0.10
			SEGRIPSHAD	-0.17	0.05	0.001	-0.26	-0.07
			SEGJANAIRT	-0.11	0.04	0.003	-0.19	-0.04
			USAVGSLOPE	0.30	0.04	0.000	0.22	0.38
			USCALCIUM	-0.13	0.04	0.003	-0.21	-0.04
			SpecMALF	0.30	0.04	0.000	0.22	0.39

Table A5.6, continued

Response variable	Type	R^2	Predictors	Estimate	S.E.	*P* - value	2.5% CI	97.5% CI
sed_pct_richness_decreaser	metric	0.512	instreamVis	-0.17	0.03	0.000	-0.24	-0.10
			CHLA	-0.14	0.03	0.000	-0.21	-0.07
			SEGRIPSHAD	-0.14	0.04	0.000	-0.21	-0.07
			SEGMINTNOR	0.13	0.04	0.000	0.06	0.20
			USAVGSLOPE	0.31	0.05	0.000	0.21	0.41
			USPHOSPHOR	0.13	0.04	0.001	0.05	0.20
			USHARDNESS	0.12	0.04	0.007	0.03	0.20
			SpecMALF	0.35	0.04	0.000	0.27	0.43
sed_pct_richness_increaser	metric	0.343	instreamVis	0.21	0.04	0.000	0.13	0.29
			ORDER_	0.23	0.05	0.000	0.13	0.34
			SEGRIPSHAD	0.14	0.05	0.012	0.03	0.25
			SEGMINTNOR	-0.15	0.04	0.000	-0.23	-0.07
			USAVGSLOPE	-0.24	0.04	0.000	-0.33	-0.16
			SEGFLOWSTA	-0.17	0.05	0.001	-0.26	-0.07
			SpecMALF	-0.19	0.05	0.000	-0.29	-0.09
chl_MCI_like	metric	0.404	CHLA	-0.30	0.04	0.000	-0.37	-0.22
			maxrateToQ50	-0.12	0.04	0.001	-0.19	-0.05
			ORDER_	-0.18	0.04	0.000	-0.25	-0.10
			DSDIST2COA	0.16	0.05	0.001	0.07	0.26
			SEGMINTNOR	0.30	0.05	0.000	0.20	0.39
			USAVGSLOPE	0.24	0.04	0.000	0.16	0.32
			SEGFLOWSTA	0.25	0.04	0.000	0.18	0.32
chl_pct_richness_decreaser	metric	0.236	CHLA	-0.29	0.04	0.000	-0.37	-0.20
			maxrateToQ50	-0.12	0.04	0.003	-0.20	-0.04
			SEGJANAIRT	-0.15	0.04	0.001	-0.23	-0.06
			SEGMINTNOR	0.19	0.04	0.000	0.11	0.26
			USAVGSLOPE	0.14	0.05	0.003	0.05	0.23
			SEGFLOWSTA	0.19	0.04	0.000	0.11	0.27
chl_pct_richness_increaser	metric	0.312	CHLA	0.27	0.04	0.000	0.19	0.34
			DSDIST2COA	-0.16	0.05	0.001	-0.26	-0.06
			SEGMINTNOR	-0.18	0.05	0.000	-0.28	-0.08
			USPHOSPHOR	0.25	0.04	0.000	0.17	0.33
			SEGFLOWSTA	-0.21	0.04	0.000	-0.29	-0.13
sed_richness_decreaser	metric	0.309	instreamVis	-0.13	0.04	0.001	-0.22	-0.05
			maxrateToQ50	-0.11	0.04	0.003	-0.19	-0.04
			ORDER_	-0.18	0.04	0.000	-0.26	-0.10
			DSDIST2COA	0.18	0.05	0.001	0.08	0.29
			SEGJANAIRT	-0.14	0.04	0.001	-0.22	-0.06
			SEGMINTNOR	0.28	0.05	0.000	0.18	0.39
			USAVGSLOPE	0.26	0.04	0.000	0.17	0.35
			SpecMALF	0.29	0.05	0.000	0.20	0.38

Table A5.6, continued

Response variable	Type	R^2	Predictors	Estimate	S.E.	*P* - value	2.5% CI	97.5% CI
sed_richness_increaser	metric	0.444	instreamVis	0.21	0.04	0.000	0.14	0.28
			DSDIST2COA	0.12	0.04	0.001	0.05	0.19
			SEGRIPSHAD	0.14	0.04	0.000	0.07	0.21
			SEGJANAIRT	-0.11	0.04	0.002	-0.18	-0.04
			USAVGSLOPE	-0.25	0.04	0.000	-0.33	-0.18
			SpecMALF	-0.29	0.04	0.000	-0.37	-0.21
chl_richness_decreaser	metric	0.249	CHLA	-0.16	0.04	0.000	-0.23	-0.08
			maxrateToQ50	-0.17	0.04	0.000	-0.25	-0.09
			ORDER_	-0.22	0.04	0.000	-0.30	-0.14
			DSDIST2COA	0.16	0.05	0.003	0.06	0.27
			SEGJANAIRT	-0.17	0.04	0.000	-0.26	-0.09
			SEGMINTNOR	0.28	0.05	0.000	0.18	0.38
			SEGFLOWSTA	0.25	0.04	0.000	0.17	0.33
chl_richness_increaser	metric	0.366	CHLA	0.33	0.04	0.000	0.26	0.41
			SEGJANAIRT	-0.29	0.04	0.000	-0.37	-0.22
			SpecMALF	-0.29	0.04	0.000	-0.36	-0.22
1b-SIZE1	trait	0.227	instreamVis	-0.17	0.04	0.000	-0.26	-0.09
			CHLA	-0.15	0.04	0.001	-0.24	-0.06
			SEGRIPSHAD	-0.14	0.04	0.002	-0.23	-0.05
			SEGJANAIRT	0.16	0.05	0.003	0.05	0.27
			USAVGSLOPE	0.22	0.06	0.000	0.11	0.34
			USPHOSPHOR	0.20	0.06	0.001	0.08	0.31
			USHARDNESS	0.17	0.05	0.001	0.07	0.27
3b-UNIV	trait	0.385	instreamVis	-0.14	0.04	0.000	-0.22	-0.07
			CHLA	-0.16	0.04	0.000	-0.24	-0.09
			maxrateToQ50	-0.15	0.04	0.000	-0.22	-0.08
			SEGRIPSHAD	-0.20	0.04	0.000	-0.27	-0.13
			SEGJANAIRT	-0.15	0.05	0.002	-0.24	-0.05
			USPHOSPHOR	-0.35	0.05	0.000	-0.45	-0.24
			SpecMALF	0.26	0.04	0.000	0.19	0.34
3c-	trait	0.408	instreamVis	0.12	0.04	0.001	0.05	0.20
			CHLA	0.14	0.04	0.000	0.07	0.22
			maxrateToQ50	0.10	0.04	0.007	0.03	0.17
			DSDIST2COA	-0.17	0.05	0.001	-0.27	-0.08
			SEGRIPSHAD	0.13	0.04	0.000	0.06	0.21
			SEGMINTNOR	-0.22	0.05	0.000	-0.32	-0.12
			USAVGSLOPE	-0.13	0.05	0.004	-0.22	-0.04
			USPHOSPHOR	0.20	0.04	0.000	0.12	0.28
			SpecMALF	-0.23	0.04	0.000	-0.32	-0.15

Table A5.6, continued

Response variable	Type	R^2	Predictors	Estimate	S.E.	*P* - value	2.5% CI	97.5% CI
6b-HERMA	trait	0.393	CHLA	0.21	0.04	0.000	0.14	0.29
			SEGJANAIRT	-0.11	0.04	0.003	-0.19	-0.04
			SEGMINTNOR	-0.12	0.04	0.001	-0.19	-0.05
			USAVGSLOPE	-0.29	0.04	0.000	-0.37	-0.21
			SpecMALF	-0.26	0.04	0.000	-0.34	-0.19
6c-TWO	trait	0.403	CHLA	-0.12	0.04	0.002	-0.19	-0.04
			SEGMINTNOR	0.14	0.03	0.000	0.07	0.21
			USAVGSLOPE	0.38	0.04	0.000	0.30	0.46
			SpecMALF	0.30	0.04	0.000	0.22	0.37
7b-SUBMERGED	trait	0.410	instreamVis	0.17	0.04	0.000	0.10	0.25
			maxrateToQ50	0.12	0.04	0.001	0.05	0.19
			ORDER_	0.17	0.04	0.000	0.10	0.25
			DSDIST2COA	-0.19	0.05	0.000	-0.29	-0.09
			SEGMINTNOR	-0.28	0.05	0.000	-0.37	-0.18
			USAVGSLOPE	-0.39	0.04	0.000	-0.47	-0.31
			SpecMALF	-0.24	0.04	0.000	-0.33	-0.16
8a-EGGFREE	trait	0.273	DIN	-0.18	0.05	0.000	-0.27	-0.09
			ORDER_	-0.19	0.04	0.000	-0.27	-0.11
			SEGMINTNOR	0.21	0.04	0.000	0.12	0.29
			USAVGSLOPE	0.49	0.05	0.000	0.40	0.59
			USPHOSPHOR	0.27	0.04	0.000	0.18	0.36
10b-CRAWLER	trait	0.301	instreamVis	-0.23	0.04	0.000	-0.31	-0.16
			SEGRIPSHAD	-0.19	0.04	0.000	-0.26	-0.11
			USAVGSLOPE	0.23	0.04	0.000	0.15	0.32
			USPHOSPHOR	-0.22	0.04	0.000	-0.30	-0.14
11b-LOWFLE	trait	0.451	CHLA	-0.23	0.04	0.000	-0.30	-0.16
			DSDIST2COA	0.14	0.05	0.003	0.05	0.24
			SEGRIPSHAD	-0.14	0.04	0.000	-0.21	-0.07
			SEGMINTNOR	0.21	0.05	0.000	0.12	0.31
			USAVGSLOPE	0.17	0.04	0.000	0.09	0.25
			USPHOSPHOR	-0.27	0.04	0.000	-0.35	-0.19
			SpecMALF	0.16	0.04	0.000	0.08	0.24
13b-SCRAPER	trait	0.136	instreamVis	-0.14	0.04	0.001	-0.23	-0.05
			DSDIST2COA	-0.20	0.04	0.000	-0.28	-0.11
			USAVGSLOPE	0.21	0.04	0.000	0.13	0.29
13d-FILTERFREE	trait	0.086	instreamVis	0.19	0.05	0.000	0.10	0.28
			ORDER_	0.18	0.05	0.000	0.09	0.27
			DSDIST2COA	0.13	0.04	0.005	0.04	0.21
			SEGJANAIRT	0.15	0.04	0.001	0.06	0.23

Appendix 6. Exploring the development of a multi-metric index of ecosystem health using an observed divided by expected (O/E) approach.

We developed random forest (RF) models to predict the reference condition for candidate metrics. For each metric, contemporary state was predicted in response to current land use and environment conditions (i.e. model predictors as outlined in Table 15 on page 71). The predictive accuracy of RF models was assessed using mean square error (i.e. the smaller the better) and coefficient of determination (i.e. the larger the better) of the relationship between observed and predicted contemporary metric values (Figure A6.1). R^2 values for the RF models ranged from 0.40 (10b-Crawler) to 0.77 (4a-CPI1) showing satisfactory to very good model performance.

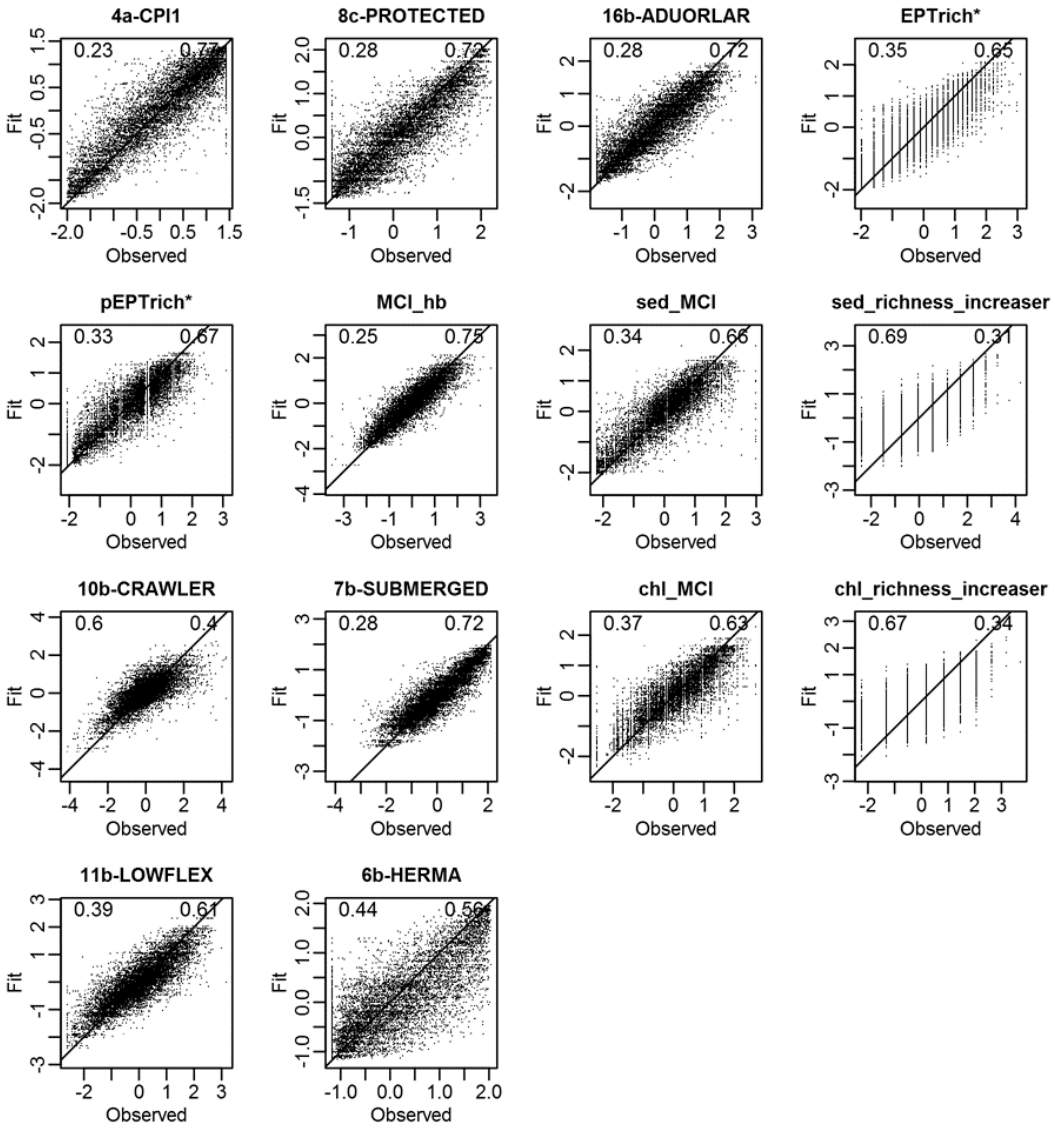


Figure A6.1 Observed versus predicted metric values from a random forest model. Scores have been normalised using Box-Cox transformation. The mean square error for each RF model is shown in the top left hand corner and the R^2 value in the top right hand corner.

Then land use was reset to natural state (e.g. native vegetation cover = 100%, urban and pastoral cover = 0%) and reference condition predicted using the same model. We plotted predicted reference condition against measured metric values to assess the range in potential O/E scores for each metric. Ideally for metrics that decrease in response to land use, the majority of E values should be above the 1:1 line (e.g. 4a-CPI1 and MCI_hb, Figure A6.2) and the opposite for metrics that increase in response to land use (e.g. 7b-Submerged, Figure A6.2). However, for many metrics this was not the case indicating that if O/E was calculated using modelled E then there would be numerous values > 1, as was the case.

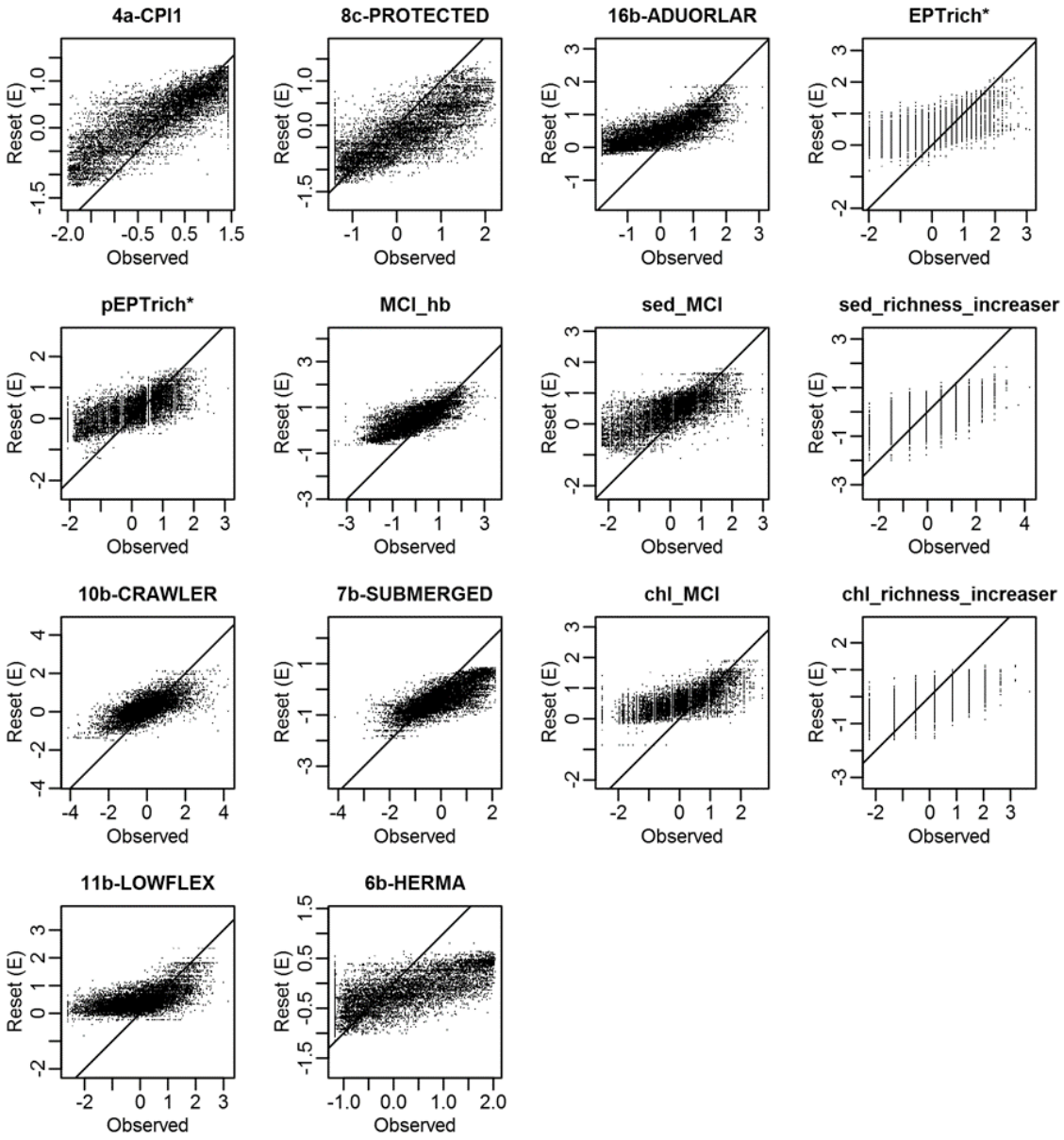


Figure A6.2. Reference condition predicted from a random forest model where land use was reset to natural state (Reset (E)) versus observed metric scores. Scores have been normalised using Box-Cox transformation.

Despite poor model performance for some metrics we applied the same procedure outlined in Section 6.3.3 for developing an MMI using O/E metric scores instead of O metric scores. For each metric, O/E scores were standardised to 0-1 by truncating all value > 1 to equal 1. Then for each EH component (tolerance, functional aspect, diversity/richness, organisation/composition) we averaged metric O/E scores and sequentially removed metrics to determine the optimum number of metrics to discriminate reference from non-reference as measured by the AUC statistic. The metrics selected were the same as for when O values were used (see Section 6.3.3), with the exception of 8c-Protected which was retained as an indicator of functional aspects and 6b-Herma retained as a Chla tolerance trait (Table A6.1). The AUC scores (ability to distinguish reference from non-reference) were lower than that observed using O values alone (e.g. 0.7 compared to 0.8 values).

Table A6.1. The AUC scores used to select the optimum combination of metrics for each ecosystem health component and the combined overall multi-metric index. * excluding Hydroptilidae.

EH component	Contributing metrics	AUC
Functional aspects	All three (4a-CPI1, 8c-Protected,16b-AduorLar)	0.75
Diversity/richness	EPT richness*	0.72
Organisation/composition	% EPT richness*	0.67
Tolerance - general	MCI_hb	0.75
Tolerance - sediment	All four (Sed_MCI, Sed_rich_increasers, 10b-Crawlers, 7b-Submerged)	0.70
	Three (Sed_MCI, 10b-Crawlers, 7b-Submerged)	0.74
Tolerance - chla	All 4 (Chla_MCI, Chla_rich_increasers, 11b-Lowflex, 6b-Herma)	0.72
	Three (Chla_MCI, 11b-Lowflex, 6b-Herma)	0.74
MMI	Weighted equally by component	0.79