



## Finding reference: a comparison of modelling approaches for predicting macroinvertebrate community index benchmarks

J. E. Clapcott, E. O. Goodwin, T. H. Snelder, K. J. Collier, M. W. Neale & S. Greenfield

To cite this article: J. E. Clapcott, E. O. Goodwin, T. H. Snelder, K. J. Collier, M. W. Neale & S. Greenfield (2017) Finding reference: a comparison of modelling approaches for predicting macroinvertebrate community index benchmarks, New Zealand Journal of Marine and Freshwater Research, 51:1, 44-59, DOI: [10.1080/00288330.2016.1265994](https://doi.org/10.1080/00288330.2016.1265994)

To link to this article: <http://dx.doi.org/10.1080/00288330.2016.1265994>



View supplementary material [↗](#)



Published online: 20 Dec 2016.



Submit your article to this journal [↗](#)



Article views: 64



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



RESEARCH ARTICLE



# Finding reference: a comparison of modelling approaches for predicting macroinvertebrate community index benchmarks

J. E. Clapcott<sup>a</sup>, E. O. Goodwin<sup>a</sup>, T. H. Snelder<sup>b</sup>, K. J. Collier<sup>c\*</sup>, M. W. Neale<sup>d†</sup> and S. Greenfield<sup>e</sup>

<sup>a</sup>Cawthron Institute, Nelson, New Zealand; <sup>b</sup>Land Water People Ltd, Lyttelton, New Zealand; <sup>c</sup>Waikato Regional Council, Hamilton, New Zealand; <sup>d</sup>Research, Investigations and Monitoring Unit, Auckland Council, Auckland, New Zealand; <sup>e</sup>Greater Wellington Regional Council, Wellington, New Zealand

## ABSTRACT

Reference benchmarks are needed to assess the contemporary status of rivers and to establish restoration targets. We developed predictive models to estimate site-specific reference values for a macroinvertebrate community index (MCI), which is used to indicate a range of human impacts on wadeable streams. We compared three statistical modelling approaches – general linear, boosted regression tree (BRT) and random forest (RF) – and tested the effect of spatial scale on predictive accuracy by developing national and regional BRT models. Using fitted flexible models (BRT, RF) and resetting predictors to reflect natural state provided the most accurate predictions of reference condition. Variation in reference MCI predictions from national and regional models was within the range observed from methodological and temporal variability. The proportion of native vegetation in upstream catchments was the primary predictor of MCI scores in all models, while secondary predictors varied regionally.

## ARTICLE HISTORY

Received 22 July 2016

Accepted 24 November 2016

## KEYWORDS

Macroinvertebrates; biotic indices; predictive models; reference condition; New Zealand; MCI


## Introduction

Reference condition provides a baseline for assessing the contemporary status of streams and rivers, and for establishing restoration benchmarks. Approaches to defining reference condition in modified landscapes include the use of (i) minimally disturbed sites, (ii) historical datasets and (iii) best available or best attainable condition (Stoddard et al. 2006). The ‘reference condition’ is commonly characterised by first stratifying natural variation using stream classifications. Sites within individual classes are then chosen that represent reference state based on the absence of anthropogenic stressors. The reference condition is then quantified for biotic or water quality measures based on surveys of the chosen reference sites. Pressure filters are used to set an acceptable level of naturalness and hence identify sites that are in a reference state, for example, sites with no introduced species present and with greater than 85% natural vegetation in their catchments

**CONTACT** J. E. Clapcott ✉ [joanne.clapcott@cawthron.org.nz](mailto:joanne.clapcott@cawthron.org.nz)

<sup>\*</sup>Current address: University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

<sup>†</sup>Current address: MartinJenkins Ltd, PO Box 7551, Auckland 1141, New Zealand.

 Supplemental data for this article can be accessed at <http://dx.doi.org/10.1080/00288330.2016.1265994>

© 2016 The Royal Society of New Zealand

(Collier et al. 2007). Stream classifications are used to control for biogeographical or ecological variability so that sites are assessed against relevant benchmarks (Hawkins et al. 2010). However, the spatial distribution of human pressures means there are inevitably some stream classes for which there are very few or no comparable reference sites, for example, lowland streams in New Zealand (Unwin et al. 2014). When reference sites are sparse or absent, models can provide an alternative method for estimating reference condition through spatial interpolation (e.g. Dodds & Oakes 2004). Indeed, site-specific modelling approaches have become the focus of defining reference benchmarks for assessing ecological condition (Hawkins et al. 2010).

Models that predict reference condition for use in stream biomonitoring are based on the assumption that a quantifiable relationship exists between anthropogenic stressor(s) and biotic metric response. A simple linear relationship between pressure and response variables enables a linear regression model to be used to estimate the reference condition. For example, McDowell et al. (2013) used analysis of covariance (ANCOVA) to predict the reference condition of water quality indicators in classes of New Zealand streams based on relationships with the proportion of catchment occupied by high producing exotic grassland. However, ecological responses are rarely linear and complex relationships are common, especially as the complexity of response variables increases from population- to community- or ecosystem-level metrics (Wu & David 2002; Clapcott et al. 2012). More flexible statistical methods, such as random forests (RF), allow for non-linear relationships and high-order interactions and are increasingly used in community and ecosystem-level modelling (Clapcott et al. 2012; Booker et al. 2014). RF and boosted regression trees (BRT) are machine learning methods that provide an advanced form of regression (Elith et al. 2008) and can provide more accurate estimates of stream metrics than other model approaches (Waite et al. 2012; Waite 2014), although the accuracy of reference condition predictions has yet to be tested.

Previous studies have suggested that large-scale models that lack predictors describing the proximate causes of variation provide little insight into causal links between land cover and stream biotic response observed at a local scale (Allan 2004; King et al. 2005; Riseng et al. 2011). Consequently regional models, where training data are grouped by biogeographic or physiographic similarity and proximate variables are measured, have been shown to provide more accurate predictions of biotic metrics (Ode et al. 2008; Cuffney et al. 2011; May et al. 2015). However, when sufficient data are available to inform the full gradient of environmental descriptors and their potential interactions, there is no reason why a broad-scale model would be less accurate than a regional model. Furthermore, combining data from multiple surveys increases information and potentially improves model performance when compared to smaller datasets.

We developed predictive models for a macroinvertebrate community index (MCI) that can be used to predict the expected reference condition of wadeable streams across New Zealand. We tested three modelling approaches – ANCOVA, RF and BRT – to assess which one(s) provided the most accurate predictions of reference condition. We expected predictions of MCI reference condition (i) would differ across different stream classes defined by climate and source of flow due to the underlying influence of hydroclimatic factors (Leathwick et al. 2011), and (ii) would be more accurate for regional models than for a national model, held to the same training  $n$ , because environmental variability is limited within regions.

## Methods

### *Benthic invertebrate data*

Benthic invertebrate metric data were compiled from regional council river monitoring databases and a national river monitoring network in New Zealand (Smith et al. 1996). This national dataset comprised 1033 sites sampled multiple times and spanning wide environmental and impact gradients across the North and South islands. We restricted our analyses to the median of MCI values from five years spanning 2007–2011, so that responses were comparable to land cover predictors derived from the most recent satellite imagery available (2007–2008; Land Cover Database 3). This dataset was supplemented by 231 sites from the Wellington and Auckland-Waikato regions sampled once over 2010–2012.

Invertebrate samples from each site were collected either quantitatively with a Surber sampler (0.1 m<sup>2</sup>) or semi-quantitatively with a kick-net (0.5 mm mesh size) of 0.6–1.0 m<sup>2</sup> of the streambed within a riffle or run habitat using standard protocols (Scarsbrook et al. 2000; Stark et al. 2001). Sampling effort can affect the number of taxa collected but has a minimal (<10%) effect on community index scores (Stark 1993). The majority of samples (>70%) were collected during summer (December–February) and were preserved in the field in c. 70% ethanol or isopropyl alcohol. Invertebrates were identified in the laboratory using standard guides (e.g. Winterbourn et al. 2006), primarily to genus for insects and most molluscs, and phylum to family for most other taxa corresponding to the level of taxonomic resolution used by Stark et al. (2001) for calculating the MCI. This tolerance metric was developed for indicating organic enrichment in wadeable streams with hard bed substrate (Stark 1985), where site scores potentially range from >150 (indicating very healthy) to as low as 20 (very poor). MCI scores were logit transformed prior to analysis because during pilot studies we observed consistently better model validation for transformed compared to non-transformed scores, and no difference in predictor variable rank importance. MCI scores were back transformed to provide meaningful values and hence we checked for, and where necessary corrected for, retransformation bias prior to model testing using the method of Manning (1998).

### *Environmental data*

We extracted a set of continuous environmental variables from the Freshwater Ecosystems of New Zealand (FENZ) database (Leathwick et al. 2010) that had been previously shown to provide reliable predictions of invertebrate distributions (Leathwick et al. 2011; Clapcott et al. 2012). Environmental variables used included catchment characteristics (catchment averages values) and segment-scale measures of geology and topography, slope, flow and flow-influencing factors (e.g. upstream rainfall) (Table 1). Upstream catchment land cover data for the downstream node of each segment were summarised from 34 categories into the proportional cover of five categories (Table 1), based on the most recent satellite imagery available (2007–2008; Land Cover Database 3). We also calculated a measure of surface water allocation (SWA) pressure as an index of the effects anthropogenic water use has on river flows. For each segment in the network, we summed the upstream consented daily water allocation (MfE 2006) and expressed it as a proportion of the mean annual low flow. The mean annual low flow for each segment was estimated using the

**Table 1.** Description of land cover and land-use stressors and other continuous environmental variables and the mean and range in values in the datasets used for national ( $n = 1264$ ) and regional models of MCI scores.

Category	Predictor	Description	National	Auckland-Waikato	Greater Wellington
<i>Land cover and land-use stressors</i>	<i>NativeVeg</i>	Native vegetation cover in the catchment (%)	34 (0,100) <sup>A,B</sup>	37.9 (0, 100)	37.8 (0, 100)
	<i>PastoralHeavy</i>	Pastoral heavy cover in the catchment (%)	43.1 (0,100) <sup>A</sup>	45.9 (0, 100)	
	<i>PastoralLight</i>	Pastoral light cover in the catchment (%)	6 (0, 92) <sup>A</sup>	0.6 (0, 40)	
	<i>Urban</i> <sup>1</sup>	Urban impervious cover in the catchment (%)	10.4 (0, 100) <sup>A</sup>	11.4 (0, 100)	
	<i>Bareground</i> <sup>2</sup>	Bare ground in the catchment (%)	0.31 (0, 7.35)		
	<i>SWA</i> <sup>3</sup>	Low flow remaining after the upstream daily water allocation is deducted (proportion)	0.25 (0, 1) <sup>A</sup>	0.11 (0, 1)	0.25 (0, 1)
<i>Environmental variables</i>	<i>Flow</i>				
	<i>SegLowFlow</i> <sup>4</sup>	Mean annual 7-day low flow (m3/s)	−3 (−8, 1.7) <sup>A</sup>	−4 (−8, 0.8)	
	<i>SegFlowStability</i>	Annual low flow/annual mean flow (ratio)	0.2 (0, 0.5) <sup>A,B</sup>	0.2 (0, 0.5)	
Temperature and shading	<i>USRainDays</i>	Days/year with rainfall in the catchment greater than 25 mm	2.4 (0.6, 4.3) <sup>A,B</sup>	2.6 (1.9, 3.5)	
	<i>SegSumT</i>	Segment summer air temperature (°C)	17.2 (12.6, 19.6) <sup>A</sup>	18 (12.8, 19.3)	16.9 (16, 17.6)
	<i>USAvgT</i>	Average air temperature (°C) in the catchment, normalised with respect to SegJanAir	−0.4 (−5.98, 1.6) <sup>A</sup>	0.02 (−2.46, 1.09)	−0.3 (−3.2, 0.8)
Geology and topography	<i>SegTSeas</i>	Segment winter air temperature (°C), normalised with respect to SegJanAir	0.5 (−4.2, 3.5) <sup>A</sup>	0.7 (−0.8, 2.1)	
	<i>SegShade</i>	Segment riparian shade (proportional)	0.3 (0, 0.8) <sup>A</sup>	0.4 (0, 0.8)	
	<i>USSlope</i>	Average slope in the catchment (°)	12.6 (0, 32) <sup>A,B</sup>	10.6 (0.2, 28.4)	14.9 (0.4, 29.4)
	<i>SegSlope</i> <sup>2</sup>	Segment slope (°)	1.7 (1, 15.2) <sup>A,B</sup>	2.2 (1, 13.7)	1.7 (1, 7.3)
	<i>SegHab</i>	Weighted average of proportional cover of local habitat using categories of: 1 = still; 2 = backwater; 3 = pool; 4 = run; 5 = riffle; 6 = rapid; 7 = cascade	4 (2.3, 5.9) <sup>A</sup>	3.9 (2.5, 4.6)	
	<i>SegSubstrate</i>	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	3.5 (1, 5.9) <sup>A,B</sup>	3.2 (1, 5.1)	3.7 (1.1, 5.1)
	<i>USCalcium</i>	Average calcium concentration of rocks in the catchment, 1 = very low to 4 = very high	1.6 (1, 4) <sup>A,B</sup>	1.5 (1, 2.9)	1.5 (1, 2.7)
	<i>USHardness</i>	Average hardness of rocks in the catchment, 1 = very low to 5 = very high	2.9 (1, 5) <sup>A</sup>	2.9 (1, 4)	3 (1, 4)
	<i>USPhosphorus</i>	Average phosphorus concentration of rocks in the catchment, 1 = very low to 5 = very high	2.2 (1, 5) <sup>A</sup>	1.8 (1, 4)	
	<i>USPeat</i>	Area of peat in upstream catchment (%)	0 (0, 1)		
	<i>USLake</i>	Area of lake in upstream catchment (%)	0 (0, 0.1)		

Note: For Auckland-Waikato  $n = 492$ . For Wellington  $n = 107$ .

Data transformations indicated where relevant: <sup>1</sup>, cube root; <sup>2</sup>, square root; <sup>3</sup>, fourth root; <sup>4</sup>, double log; <sup>5</sup>, log.

Variables indicated when used in: <sup>A</sup>, restricted-national model A; <sup>B</sup>, restricted-national model B.

methods of Pearson (1995). Approximately 6.5% of stream segments at a national scale were affected by surface water takes. Of the resulting 22 predictors, 16 were approximately normally distributed. To approximate normal distributions for use in a linear model (ANCOVA) we square root transformed the *Bareground* and *SegSlope* variables, and respectively cube root, fourth root, log and double log transformed the variables *Urban*, *SWA*, *USRainDays* and *SegLowFlow*. Additionally, sites were assigned to one of 23 Climate and Source-of-Flow (CSOF) categories from the River Environment Classification (REC; Snelder & Biggs 2002). The categories represented in the dataset included Climate: warm-extremely wet (WX), warm-wet (WW), warm-dry (WD), cool-extremely wet (CX), cool-wet (CW), cool-dry (CD) and Source-of-Flow: glacial mountain (GM), mountain (M), hill (H), low elevation (L), lake (Lk).

### Regression models

We fitted three types of regression models to the MCI scores: ANCOVA, RF and BRT. ANCOVA includes categorical and continuous variables in a linear regression model (Dodds & Oakes 2004). We fitted ANCOVA models that included CSOF as the categorical variable and six continuous variables representing the proportion of the catchment occupied by various land cover and the water allocation pressure variable. We used stepwise selection to reduce the model terms to only include those variables that contributed to the predictive performance of the ANCOVA model. The approach reduces the probability that land cover may be correlated with the environmental variables because the relationship between land cover and the indices is defined for a group for which segment- and catchment-scale environmental variables are considered homogeneous. However, the method can only be applied to classes for which there is sufficient replication to define the regression relationships and does not allow estimates of reference values in non-represented classes. ANCOVA analyses were carried out by using *anova* and *lm* commands in R (R Core Team 2014).

RF models are an ensemble of individual classification and regression trees (a forest). Each tree is grown with a random bootstrap sample of the entire data set which then becomes the training data for that tree. Predictions are made by averaging individual predictions from the ensemble of trees. The structures of RF models can be examined by using importance measures and partial dependence plots. Importance measures indicate the contribution of the predictors to model accuracy. Partial dependence plots show the marginal contribution of a predictor to the response (i.e. the response as a function of the predictor when the other predictors are held at their mean value). Our RF regression models included continuous land cover and environmental variables, were trained using 500 trees, with seven variables tried at each split. We used the *randomForest* library in R following methods outlined by Breiman (2001).

BRT models also automatically fit non-linear and high-order interactions (Friedman 2001; Elith et al. 2008; Hastie et al. 2009). The BRT method combines additive regression modelling with boosting techniques, and provides an estimate from numerous and often thousands of very simple models. BRT models include a measure of the comparative strength of association between the response variable and predictor variables (percentage deviance explained) as well as a cross-validation coefficient (CV) that indicates the predictive performance of the model. BRT analysis also indicates the form of the predictor-

relationship (e.g. linear, curvilinear or sigmoidal fitted functions), which are comparable to partial dependence plots in RF models. In our BRT models, we used continuous land cover and environmental variables, tree complexity (the number of regression tree rules or splits associated with each individual model) was set at seven (owing to the high number of explanatory variables that could interact), and the learning rate was tuned to ensure at least 1000 trees were included in the final model (Elith et al. 2008). For BRT analyses, we used the *gbm* library of Ridgeway (2006) supplemented by scripts from Elith et al. (2008).

### **National models**

We used the ANCOVA, RF and BRT models to predict reference conditions at a national scale by modelling MCI as a function of environmental and human impact variables. We trained models using the entire dataset ( $n = 1264$ ) to compare model diagnostics and the relative importance of predictor variables. Then we performed a leave-one-out cross-validation (LOOCV) procedure (Hastie et al. 2009) to test the predictive performance of the models. We plotted the observations versus predictions and quantified the model performance using the following metrics; Nash–Sutcliffe efficiency (NSE), bias, and the root mean squared deviation (RMSD) (Piñeiro et al. 2008). The NSE statistic indicates how closely the observations coincide with predictions (Nash & Sutcliffe 1970). NSE values range from  $-\infty$  to 1. An NSE of 1 corresponds to a perfect match and values greater than 0.5 indicate good model performance. Bias measures the average tendency of the predicted values to be larger or smaller than the observed, where positive values indicate model underestimation and negative values indicate overestimation bias. The RMSD is an estimate of model accuracy, where smaller values indicate greater accuracy (Piñeiro et al. 2008). We used RMSD to estimate model prediction intervals. We further assessed model precision by testing whether the slope of the line of best fit was significantly different to the 1.

For both RF and BRT methods, we estimated reference conditions by first setting the values of the predictors that represented anthropogenic stressors to zero. We then predicted the value of MCI for all segments of the river network. Similarly for the ANCOVA model, we calculated reference MCI scores by assigning zero to anthropogenic variables in the regression equation for each CSOF category. We used a second LOOCV procedure to assess the relative accuracy of MCI reference predictions from the three modelling approaches. In this second LOOCV procedure, the hold-out data were restricted to minimally disturbed sites (i.e. >85% native cover, <15% light pastoral cover, <5% heavy pastoral cover and urban and water allocation = 0). We relaxed these land cover rules to ensure there were sufficient sites ( $n = 90$ ) to test the models.

### **Regional models**

To test our hypothesis that regional models would provide more accurate predictions than national models, we developed BRT models for two case study regions: Auckland-Waikato ( $n = 492$ ) and Wellington ( $n = 107$ ). For the Auckland-Waikato model, variables were excluded if they did not explain any deviance in the MCI data. For the Wellington model, we used stepwise selection based on Akaike Information Criterion to select a set



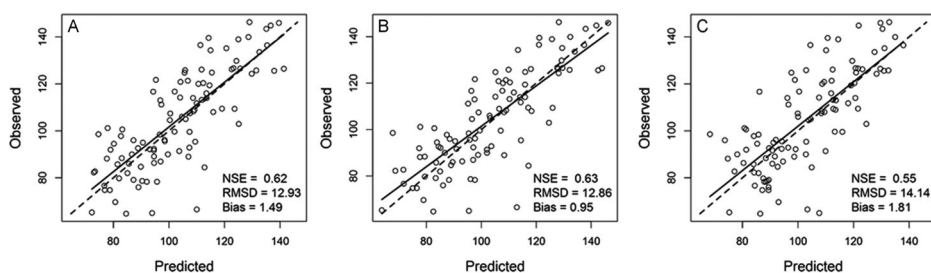
of predictor variables. Because sample size can affect the predictive performance of BRT models (Elith et al. 2008), we also developed two models using a random selection of sites from the national data set to match the  $n$  of regional models; restricted-national model A ( $n = 492$ ) and restricted-national model B ( $n = 107$ ). Variable selection was the same as for the regional models, with variables excluded from model A if they did not explain any deviance in the MCI data and stepwise selection used to select predictor variables in model B. We tested the performance of all four models using the LOOCV procedure, and compared reference MCI predictions made from the regional, national-restricted and national models with observed values at the nominated reference sites in the Auckland-Waikato ( $n = 54$ ) and Wellington ( $n = 8$ ) regions. All analyses were conducted in R version 3.1.1 (R Core Team 2014).

## Results

### National models

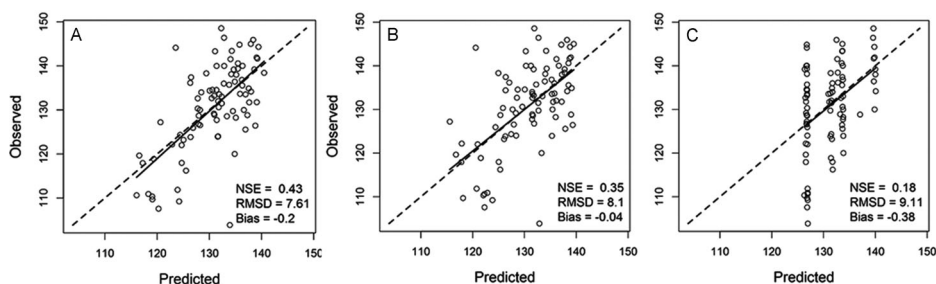
RF and BRT models provided more accurate predictions of contemporary MCI scores at a national scale compared to the ANCOVA model (Figure 1). The NSE was 0.62 and 0.63 for the RF and BRT models respectively, compared with 0.55 for the linear model. The RMSD was 12.9 for both the RF and BRT models, compared with 14.1 for the linear model. The 95% prediction interval estimated from the RMSD was 25 MCI units for both the RF and BRT models and 28 MCI units for the linear model. All model predictions had low bias, on average underestimating MCI scores by 0.95% to 1.81%. The BRT model showed the least bias (Figure 1) and the slope of the line of best fit was not significantly different to 1 in any model (data not shown).

Predicted reference MCI scores were more accurate for the RF and BRT models than the ANCOVA model. The LOOCV analysis of observed data from 90 national reference sites indicated that the RMSD was lower for the RF and BRT models, 7.6 and 8.1 respectively, compared with 9.1 for the ANCOVA model (Figure 2). Bias was less than 1% for all models and the slope of the line of best fit was not significantly different to 1 in any model (data not shown). NSE for the RF and BRT models was 0.43 and 0.35 respectively, compared with 0.18 for the linear model (Figure 2), and were much lower for reference state models than contemporary models (see Figure 1). However, absolute uncertainty was lower for the reference state predictions than for the contemporary MCI predictions.



**Figure 1.** Scatter plots of observed versus predicted values from a, RF, b, BRT and c, linear (ANCOVA) model of MCI scores at 100 randomly selected national sites. The dashed line is the 1:1 line and the solid line is the line of best fit. Model performance statistics are explained in the text.

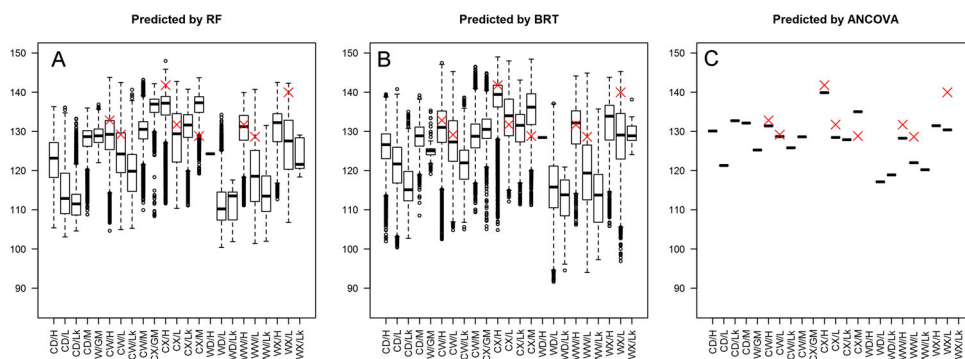




**Figure 2.** Scatter plots of observed versus predicted values from refitted a, RF, b, BRT and c, linear (ANCOVA) model of MCI scores at 90 national reference sites. The dashed line is the 1:1 line and the solid line is the line of best fit. Model performance statistics are explained in the text.

For example, the 95% prediction interval for BRT reference model predictions was 16 MCI units compared to 25 MCI units for the BRT contemporary model. The predicted mean reference MCI scores from the national models were all similar; RF = 129, BRT = 131 and ANCOVA = 128.

The LOOCV highlighted that the ANCOVA model predicts a single reference MCI score for each CSOF class whereas observations indicated that there is a range of reference MCI scores in classes (Figures 2 and 3). Native vegetation cover was the most important land cover predictor of MCI scores in all three national models (Table 2). For the RF and BRT models, the main environmental predictors of MCI scores included flow variability (*SegFlowStability*), habitat category (*SegHab*), substrate composition (*SegSubstrate*), summer temperature (*SegSumT*) and average upstream slope (*USSlope*) (Table 2). The MCI scores had monotonic responses to *NativeVeg* and *PastoralHeavy* in RF and BRT models and a strong positive response to *NativeVeg* and negative response to *PastoralHeavy* in the ANCOVA model. There were no interactions between *NativeVeg*, *PastoralHeavy* and CSOF retained in the ANCOVA model, meaning that all classes showed similar slopes for the relationships with those land cover predictors.



**Figure 3.** Summary boxplots (median and interquartile range, 1.5x interquartile range and outliers) of predicted reference MCI scores for CSOF classes at a national scale from a, RF, b, BRT and c, linear (ANCOVA) models. Crosses show the median MCI value observed at reference sites. Summary values are provided in Table S1.

**Table 2.** Comparison of predictor variable relative importance (sums-of-squares for ANCOVA) and summary output of three approaches to modelling MCI at the national and regional scales.

Category	Predictor	National BRT	National RF	National ANCOVA	Auckland-Waikato BRT	Wellington BRT
<i>Land cover and land-use stressors</i>	<i>Bareground</i>	0.36	0.39			
	<i>NativeVeg</i>	<b>33.71</b>	<b>20.26</b>	43.04	<b>26.99</b>	<b>36.16</b>
	<i>Urban</i>	4.63	3.46	6.41	2.25	
	<i>PastoralHeavy</i>	<b>6.18</b>	<b>10.64</b>	3.65	3.87	
	<i>PastoralLight</i>	0.39	0.86	0.28	0.12	
<i>Environmental variables</i>	<i>SWA</i>	0.85	1.61	1.32	0.24	5.87
	<i>Climate/Source-of-Flow</i>			69.73		
	<i>Flow</i>					
	<i>SegLowFlow</i>	1.69	2.57		1.89	
	<i>SegFlowStability</i>	7.32	6.54		4.92	
Temperature and shading	<i>USRainDays</i>	4.78	4.83		3.50	
	<i>SegSumT</i>	<b>7.96</b>	4.92		<b>15.41</b>	4.88
	<i>USAvgT</i>	1.83	2.64		1.04	5.10
	<i>SegTSeas</i>	2.94	2.51		1.36	
	<i>SegShade</i>	3.71	3.71		6.74	
Geology and topography	<i>USSlope</i>	4.61	<b>8.30</b>		10.64	<b>10.32</b>
	<i>SegSlope</i>	3.56	3.82		1.21	9.70
	<i>SegHab</i>	4.63	6.32		0.79	
	<i>SegSubstrate</i>	3.75	7.71		<b>11.37</b>	5.05
	<i>USCalcium</i>	2.44	2.78		2.82	<b>17.64</b>
	<i>USHardness</i>	2.21	3.34		2.79	5.29
	<i>USPhosphorus</i>	2.17	2.40		2.05	
	<i>USPeat</i>	0.17	0.27			
	<i>USLake</i>	0.10	0.13			
	Contemporary mean MCI	104	104	103	101	105
Reference mean MCI		131	129	128	123	122

Note: The three most important variables in models are highlighted in bold text. See [Table 1](#) for a description of predictor variables and text for description of modelling approaches.

### Regional models

Model diagnostics indicated good performance of the regional BRT models for the Auckland-Waikato and Wellington regions. For the Auckland-Waikato region, the national and restricted-national model A performed better for predicting reference MCI scores than the regional model based on LOOCV at 54 sites ([Table 3](#)). NSE was higher for both national models than the Auckland-Waikato regional model. RMSD was also lower for the national and restricted-national models than the regional Auckland-Waikato model ([Table 3](#)). All three models had considerable negative bias, slopes deviated significantly from 1, and reference MCI scores in the Auckland-Waikato region were over-estimated between 7.5% and 9.7% ([Table 3](#)).

For the Wellington region, the relationships between predicted and observed reference MCI scores (from the LOOCV at eight sites) suggested that the regional model performed similar to the national model ( $NSE > 0.5$ ) with equivalent error (RMSD; [Table 3](#)). The regional model was biased and tended to underestimate reference MCI scores, albeit less so than the national model which overestimated scores. The restricted-national model B had similar model performance, but with less bias, less precision and higher error ([Table 3](#)).

**Table 3.** Comparison of performance metrics for leave-one-out cross-validations of three BRT models predicting reference MCI scores for two regional datasets.

Model	Training <i>n</i>	Test <i>n</i>	NSE	Bias	PSlope	RMSD
Auckland-Waikato	492	54	0.26	−9.74	<0.001	17.53
National	1264	54	0.38	−9.52	<0.001	16.06
Restricted-national A	492	54	0.37	−7.47	<0.001	16.21
Wellington	107	8	0.54	2.35	n.s.	16.81
National	1264	8	0.54	−5.15	n.s.	16.80
Restricted-national B	107	8	0.50	−0.64	n.s.	17.49

Note: Training *n* = number of sites used in model development, Test *n* = number of sites used in validation dataset. NSE = Nash–Sutcliffe efficiency; RMSD = root mean squared deviation. PSlope = probability that the slope of the line of best fit is significantly different to 1.

Native vegetation was the most important land cover predictor at a regional scale (Table 2). SWA was also retained as an important predictor in the Wellington regional model. Environmental variables that explained the most variance in MCI scores included summer temperature (*SegSumT*) and substrate composition (*SegSubstrate*) in the Auckland-Waikato model, and average upstream slope (*USSlope*) and calcium concentration of rocks in the catchment (*USCalcium*) in the Wellington model.

## Discussion

### *Which model provides the most accurate predictions of MCI reference condition?*

Our results show that reference MCI scores can be accurately predicted by regression models and that more flexible models perform better than a linear model. Other studies that have used flexible modelling approaches to predict biological indices of stream health have shown similar model performance (Booker et al. 2014; Pilière et al. 2014; Waite 2014). However, there are few examples of independent tests of contemporary or reference state predictions in the literature. For example, Booker et al. (2014) used a LOOCV technique similar to our study to test contemporary MCI predictions and their NSE measure was 0.64 (comparable NSE in our study = 0.62). Our independent validation statistics provide evidence of the improved predictive capability of flexible models compared to the linear (ANCOVA) model for reference state predictions.

We expected our reference predictions would overestimate MCI scores for minimally disturbed sites given that these sites have some degree of anthropogenic impact. However, we observed the opposite and this may be because BRT and RF models do not extrapolate beyond the range of the observations in the training data set (due to the recursive partitioning approach of both models). As such, sufficient reference sites across a range of environmental conditions are needed to predict reference state using this model approach. In contrast, the ANCOVA model will predict outside the range of training data observations and while this could lead to inappropriate out-of-range response predictions (e.g. MCI greater than 200) the linear model also resulted in underestimations of reference MCI scores in our study. Underestimation was consistent across the full range of MCI scores (i.e. slope was not significantly different to one). However, bias of all national models was small, being less than 2% for flexible models and less than 4% for the linear model. These percentages relate directly to MCI scores, hence a

2% bias means the BRT model predicts a mean reference MCI score of 129 instead of the observed 132.

Achieving good performance from machine learning methods generally requires including a suite of predictor variables. For example, our RF and BRT models included predictors that characterised catchment climate, topography and land cover (Table 2). Land use is generally highly correlated with the other catchment characteristics. Some authors consider this can make it difficult to use these models to isolate the contribution of land use or to predict the effect of changing land use or management (e.g. Oehler & Elliott 2011). However, we showed that setting anthropogenic stressors to zero in our fitted BRT and RF models produced accurate estimates of reference condition MCI. This is because models fitted with tree-based methods such as RF and BRT are robust to collinearity among predictors (Elith et al. 2008); the most informative predictor is chosen in each subsequent node fit so the influence of correlated predictors is minimised. When it comes to apportioning variance such as in the reference model approach, variance is primarily assigned to native vegetation cover as the most informative predictor. This could lead to less spatial variation in predicted MCI reference scores than observed, however this was not the case in our study. Numerous studies have shown native vegetation cover to be the most important predictor of MCI reference scores at multiple spatial scales (Death & Collier 2010; Clapcott et al. 2012). Furthermore, Clapcott et al. (2014) showed that the difference in variance explained in MCI scores predicted from models fitted with and without land-use predictors was equivalent to the variance assigned to land use in an 'all-in' model. This suggests that it is reasonable to expect robust reference state predictions of MCI scores from flexible models which reset land cover to 100% native vegetation.

### ***Accounting for sources of variability***

There are several sources of uncertainty in the prediction of MCI reference scores that are important to quantify when assessing the efficacy of predictive models. First, within-site spatial variability is not accounted for in our models because we used mean site scores. Stark (1993) found within-site variability using kick-net and Surber sampling methods to be in the range 10–15% of the mean value, suggesting that within-site spatial variability was not a major source of variation in model predictions. Second, seasonal and inter-annual variability are not accounted for in our models. Previous studies have shown that the seasonal effect on MCI scores is relatively small, with most seasonal means within 3% of annual means (Stark & Phillips 2009). Inter-annual variation was also relatively small (4–8%) at reference sites sampled over 10 years in the Waikato region (Collier 2008a, 2008b). Our national dataset was dominated by samples collected in summer (70%) and restricted to 5 years of consecutive data; conservatively, we estimate the additive effect of natural temporal variation on reference MCI scores to be less than 10%. Third, natural among-site variation is accounted for in our models and is shown by the range in MCI scores predicted nationally or within stream classes. Predicted MCI reference scores range from 112 to 146 at the national level, although the effective range is much less for specific CSOF classes. For example, the BRT model predicted a 5th–95th percentile range of 119–132 for sites grouped post hoc into the CD/H class and 109–131 for the CD/L class (Table S1). Variation in MCI reference scores among classes was expected

due to the range of environmental variability among classes, represented as a categorical variable in the ANCOVA model (e.g. *CSOF*) and continuous variables in the flexible models (e.g. *SegSumT*, *USSlope*). In summary, model predictions of MCI reference scores are relatively accurate when compared to the error that can be expected due to sampling error and temporal variability. Environmental variability contributes to broad-scale variation in reference predictions, but native vegetation cover accounts for a third of the variation in MCI scores (Table 2).

### ***Does a regional model provide more accurate predictions than a national model?***

Our results did not support the hypothesis that regional models would provide more accurate predictions of reference MCI scores than a national model due to better representation of locally important variables in the regional models. Instead, the efficacy of developing a regional model appears context dependent, influenced in part by the environmental variability observed within a region compared to the variability observed at the national scale. Using BRT models of stream macroinvertebrate indices, Waite et al. (2014) observed that regional models of a narrow extent provided more accurate predictions, when locally relevant environmental variables were retained, than broader-scale models. In our study, where national models were just as informative as regional models, the predictor set was not changed to reflect regional character. Furthermore, there was less signal (i.e. index response to pressure gradients) to noise (i.e. index response to environmental variability) in the Auckland-Waikato region, despite large training and validation datasets, which may further account for the lack of improvement provided by the regional model.

Good model performance for the regional model and restricted-national model for the Wellington region suggests that training datasets could be as small as 100 sites to obtain accurate predictions. While model validation statistics can be strongly influenced by sample size (Hastie et al. 2009), our results suggest that  $n$  alone is not limiting model performance, for example, low RMSD when validating the Wellington model at eight sites. Instead it may be that the range in MCI scores observed at regional reference sites is limited and biasing validation. Inspection of the data reveals that regional reference sites are subject to very little additional land-use pressures, on average less than 1% heavy pastoral cover, in comparison to national reference sites, where pastoral cover ranges from 0% to 14%. Furthermore, regional reference sites were ground-truthed prior to selection (Collier et al. 2007), so it is unlikely that unmeasured pressures such as road crossings or point sources are causing lower observed than predicted MCI scores at reference sites.

### ***Management implications***

Our results show that high MCI scores occur at sites with high native vegetation in the upstream catchment, lower temperatures, larger substrate sizes, greater segment slopes and more stable flows. The RF and BRT models provide site-specific predictions which can then be grouped *post hoc* by any stream classification (Figure 3, Table S1) whereas the ANCOVA model predicts average values for *a priori* defined stream classes. As

such, a major difference between flexible and linear modelling approaches for predicting reference site metrics is the ability of the former to predict to stream classes that were not represented in the data set but have similar environmental gradients to classes that were represented. Furthermore, site-specific predictions identify a range in reference conditions and this suggests that 'reference condition' should be described as a distribution rather than a single endpoint (Stoddard et al. 2006).

The RF and BRT models provide accurate predictions of MCI scores that can be used to inform management benchmarks. Currently, four quality classes are used to denote 'Excellent' (MCI >119), 'Good' (100–119), 'Fair' (80–99) or 'Poor' (<80) conditions indicative of different levels of degradation (Stark & Maxted 2007). We suggest these MCI benchmarks should not be universally applied and benchmarks specific to particular stream types are more appropriate. For example, all national models predict average reference MCI scores of 100–119 rather than >119 for WD/L and WD/Lk river classes. MCI scores from BRT model predictions (Clapcott et al. 2014) have previously been used to inform unitary and regional plans in New Zealand. For example, the Greater Wellington Regional Council Proposed Natural Resources Plan (2015) included MCI predictions to establish ecosystem health objectives for different river classes.

Improved model performance will assist the broader application of MCI predictions for freshwater management in New Zealand. Model improvement may be possible with the use of more accurate measures of predictor variables from scales relevant to those at which benthic invertebrate samples are collected (e.g. reach-scale), instead of modelled values reflecting characteristics of stream segments which can range from hundreds to thousands of metres in length. For example, robust measures of substrate size and nutrient concentrations (e.g. Clapcott & Goodwin 2014; Wagenhoff et al. *in press*) along with improved accounting of flow and temperature condition (e.g. Booker et al. 2014), may improve the predictive performance of a national MCI model.

## Acknowledgements

John Leathwick presented a paper at the 2007 New Zealand Society for Freshwater Science Conference entitled 'Reference values for MCI - what should they be?' and his boosted regression modelling of MCI was an inspiration for this study. We thank Ministry for the Environment for the collation of MCI data from regional councils and NIWA's national river water quality network. We thank Auckland Council, Waikato Regional Council and Greater Wellington Regional Council for the provision of additional MCI data. We also thank two anonymous reviewers for constructive comments on our manuscript. Guest editor: Dr John Quinn.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

Manuscript preparation was funded in part by Ministry of Business, Innovation and Employment Contract C01X1002: Aquatic Rehabilitation (Joanne Clapcott and Eric Goodwin), Waikato Regional Council (Kevin Collier) and Auckland Council (Martin Neale). Ton Snelder was funded by Ministry of Business, Innovation and Employment Contract Number C01X1318: Ngā Kete o Te Wānanga: Mātauranga.



## References

- Allan JD. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annual Review of Ecology, Evolution and Systematics*. 35:257–284.
- Booker DJ, Snelder TH, Greenwood MJ, Crow SK. 2014. Relationships between invertebrate communities and both hydrological regime and other environmental factors across New Zealand's rivers. *Ecohydrology*. 8:13–32.
- Breiman L. 2001. Random forests. *Machine Learning*. 45:5–32.
- 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*. 57:74–90.
- Clapcott J, Goodwin E. 2014. Relationships between macroinvertebrate community index and environmental drivers. Prepared for Ministry for the Environment. Cawthron Report No. 2504. 21p. Available from: <http://www.mfe.govt.nz/publications/fresh-water/relationships-between-macroinvertebrate-community-index-and-environmental>.
- 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. doi:10.1051/kmae/2014027.
- Collier KJ. 2008a. Temporal patterns in the stability, persistence and condition of stream macroinvertebrate communities: relationships with catchment land-use and regional climate. *Freshwater Biology*. 53:603–616.
- Collier KJ. 2008b. 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, Haigh A, Kelly J. 2007. Coupling GIS and multivariate approaches to reference site selection for wadeable stream monitoring. *Environmental Monitoring and Assessment*. 127:29–45.
- Cuffney TF, Kashuba R, Qian SS, Alameddine I, Cha Y, Lee B, Coles JF, McMahon G. 2011. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. *Journal of the North American Benthological Society*. 30:797–819.
- Death RG, Collier KJ. 2010. Measuring stream macroinvertebrate responses to gradients of vegetation cover: when is enough enough? *Freshwater Biology*. 55:1447–1464.
- Dodds WK, Oakes RM. 2004. A technique for establishing reference nutrient concentrations across watersheds affected by humans. *Limnology and Oceanography: Methods*. 2:333–341.
- Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*. 77:802–813.
- Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 29:1189–1232.
- Hastie T, Tibshirani R, Friedman J. 2009. *Elements of statistical learning: data mining, inference and prediction*. 2nd ed. New York: Springer; p. 745.
- Hawkins CP, Olson JR, Hill RA. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. *Journal of the North American Benthological Society*. 29:312–343.
- King RS, Baker ME, Whigham DF, Weller DE, Jordan TE, Kazyak PF, Hurd MK. 2005. Spatial considerations for linking watershed land cover to ecological indicators in streams. *Ecological Applications*. 15:137–153.
- Leathwick JR, Snelder T, Chadderton WL, 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 JR, West D, Gerbeaux P, Kelly D, Robertson H, Brown D, Chadderton WL, Ausseil A-G. 2010. Freshwater Ecosystems of New Zealand (FENZ) Geodatabase. Available from: [www.doc.govt.nz/conservation/land-and-freshwater/freshwater/freshwater-ecosystems-of-new-zealand/](http://www.doc.govt.nz/conservation/land-and-freshwater/freshwater/freshwater-ecosystems-of-new-zealand/).
- Manning WG. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*. 17:283–295.
- May JT, Brown LR, Rehn AC, Waite IR, Ode PR, Mazor RD, Schiff KC. 2015. Correspondence of biological condition models of California streams at statewide and regional scales. *Environmental Monitoring and Assessment*. 187. doi:10.1007/s10661-014-4086-x.



- 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*. 64:387–400.
- MfE 2006. Snapshot of water allocation in New Zealand. Ministry for the Environment, Wellington. ME number 782. 64p. Available from: <http://www.mfe.govt.nz/publications/fresh-water-environmental-reporting/snapshot-water-allocation-new-zealand>.
- Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology (NZ)*. 10:282–290.
- Oehler F, Elliott AH. 2011. Predicting stream N and P concentrations from loads and catchment characteristics at regional scale: a concentration ratio method. *Science of the Total Environment*. 409:5392–5402.
- Ode PR, Hawkins CP, Mazor RD. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. *Journal of the North American Benthological Society*. 27:967–985.
- Pearson CP. 1995. Regional frequency analysis of low flows in New Zealand rivers. *Journal of Hydrology (NZ)*. 33:94–117.
- Pilière A, Schipper AM, Breure TM, Posthuma L, Zwart DD, Dyer SD, Huijbregts MAJ. 2014. Unraveling the relationships between freshwater invertebrate assemblages and interacting environmental factors. *Freshwater Science*. 33:1148–1158.
- Piñeiro G, Perelman S, Guerschman J, Paruelo J. 2008. How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecological Modelling*. 216:316–322.
- R Core Team. 2014. R: a language and environment for statistical computing. Version 3.1.1. Vienna, Austria. Available from: <http://www.R-project.org>, R Foundation for Statistical Computing.
- Ridgeway G. 2006. Generalised boosted regression models. Documentation on the R Package ‘gbm’. Version 1.5-7. Available from: <http://www.i-pensieri.com/gregr/gbm.shtml> (accessed March 2008).
- Riseng CM, Wiley MJ, Black RW, Munn MD. 2011. Impacts of agricultural land use on biological integrity: a causal analysis. *Ecological Applications*. 21:3128–3146.
- Scarsbrook MR, Boothroyd IKD, Quinn JM. 2000. New Zealand’s national river water quality network: long-term trends in macroinvertebrate communities. *New Zealand Journal of Marine and Freshwater Research*. 34:289–302.
- Smith DG, McBride GB, Bryers GG, Wisse J, Mink DFJ. 1996. Trends in New Zealand’s river water quality network. *New Zealand Journal of Marine and Freshwater Research*. 30:485–500.
- Snelder TH, Biggs BJF. 2002. Multiscale river environment classification for water resource management. *Journal of the American Water Resources Association*. 38:1–15.
- Stark JD. 1985. A macroinvertebrate community index of water quality for stony streams. *Water and Soil Miscellaneous Publication*; p. 53.
- Stark JD. 1993. Performance of the Macroinvertebrate Community Index: effects of sampling method, sample replication, water depth, current velocity, and substratum on index values. *New Zealand Journal of Marine and Freshwater Research*. 27:463–478.
- Stark JD, Boothroyd IKG, Harding JS, Maxted JR, Scarsbrook MR. 2001. Protocols for sampling macroinvertebrates in Wadeable streams. Sustainable Management Fund Project No. 5103. 57p. Available from: <http://www.mfe.govt.nz/publications/fresh-water-environmental-reporting/protocols-sampling-macroinvertebrates-wadeable>.
- Stark JD, Maxted JR. 2007. A user guide for the macroinvertebrate community index. Prepared for the Ministry for the Environment. Cawthron Report No.1166. 58p. Available from: <http://www.mfe.govt.nz/publications/freshwater-publications/user-guide-macroinvertebrate-community-index>.
- Stark JD, Phillips N. 2009. Seasonal variability in the Macroinvertebrate Community Index: are seasonal correction factors required? *New Zealand Journal of Marine and Freshwater Research*. 43:867–882.
- Stoddard JL, Larsen DP, Hawkins CP, Johnson RK, Norris RH. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecological Applications*. 16:1267–1276.

- Unwin M, Larned S, Sykes J. 2014. Recommendations for new sites to improve representativeness in the New Zealand river environmental monitoring network. Prepared for Ministry for the Environment. NIWA Report No: CHC2014-054. Available from: <http://www.mfe.govt.nz/sites/default/files/media/Fresh%20water/recommendations-for-new-river-environmental-monitoring-sites-final.pdf>.
- Wagenhoff A, Clapcott JE, Goodwin EO, Young RG. [in press](#). Thresholds in ecosystem structural and functional responses to multiple stressors can inform limit setting in streams. *Freshwater Science*.
- Waite IR. 2014. Agricultural disturbance response models for invertebrate and algal metrics from streams at two spatial scales within the U.S. *Hydrobiologia*. 726:285–303.
- Waite IR, Kennen JG, May JT, Brown LR, Cuffney TF, Jones KA, Orlando JL. 2012. Comparison of stream invertebrate response models for bioassessment metrics. *Journal of the American Water Resources Association*. 48:570–583.
- Waite IR, Kennen JG, May JT, Brown LR, Cuffney TF, Jones KA, Orlando JL. 2014. Stream macro-invertebrate response models for bioassessment metrics: addressing the issue of spatial scale. *PLoS ONE* 9(3):e90944.
- Winterbourn MJ, Gregson KLD, Dolphin CH. 2006. Guide to the aquatic insects of New Zealand. *Bulletin of the Entomological Society of New Zealand*. 14:1–108.
- Wu J, David JL. 2002. A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications. *Ecological Modelling*. 153:7–26.