Environment Waikato Technical Report 2006/07

A Predictive Model of Fish Distribution and Index of Biotic Integrity (IBI) for Wadeable Streams in the Waikato Region

Prepared by:
Mike Joy
Ecology Group &
Centre for Freshwater Ecosystem Modeling and Management
Institute of Natural Resources
Massey University

For:
Environment Waikato
PO Box 4010
HAMILTON EAST

ISSN: 1172-4005

May 2005

Document #: 1058758v2

Environment Waikato
REGIONAL COUNCIL

Peer reviewed by:
Dr Kevin Collier Initials  Date 3 March 2006

Approved for release by:
Dr Vivienne Smith Initials  Date 5 March 2006

Table of Contents

Executive Summary iii

1 Introduction 1
1.1 Background 1
1.2 Using artificial neural networks for spatial modelling 1

2 Methods 2
2.1 Predictive model construction 2
2.2 Data sources 2
2.2.1 Fish 2
2.2.2 Habitat 7
2.3 Model architecture and number of variables 8
2.4 Validation with independent data 8
2.5 Model evaluation 8
2.6 Quantifying predictor variable contributions 8
2.7 Sensitivity analysis 9
2.8 Predictive IBI 9

3 Results 9
3.1 Network architecture and variables 9
3.2 Model evaluation 9
3.2.1 Species comparison 9
3.2.2 Assemblage comparison 11
3.3 Predictor variable importance 11
3.4 Sensitivity analysis 13

4 Discussion 16
4.1 Assemblage-environment relationships 16
4.2 Species-environment relationships 18
4.3 Limitations of the predictive model 18
4.4 Future data requirements 18

References 19

Appendix 1: Technical details on model construction validation and evaluation from Joy & Death (2004). 21

Appendix 2: FWENZ variables used in model construction. 23

List of Figures

Figure 1: Flow chart for predictive IBI development. 3
Figure 2: Map showing the predicted IBI scores over the Waikato Region for streams of 4th order and less. 4
Figure 3: Proportion of the 1967 sites in the Waikato region at which each of the 13 modelled species were found. 6
Figure 4: Site map showing the Waikato sample sites from the New Zealand Freshwater Fish Database. 7
Figure 5: Crossvalidated Cohen's kappa results for the 13 species see text for details on classifications. 10
Figure 6: Histogram of crossvalidated simple matching coefficients for the match between observed and predicted fish assemblages using the ANN model at the 1967 Waikato sites. 11
Figure 7: Sensitivity analysis for catchment phosphorus versus probability of occurrence. In each plot (one for each taxon) all other variables were set to their mean values and the catchment proportions of the variable were varied over its full range. Note the y-axis scales differ depending on prevalence of the taxa. 14

Figure 8: Sensitivity analysis for the proportion of the stream segment in pastoral farming versus probability of occurrence. In each plot (one for each taxon) all other variables were set to their mean values and the catchment proportions of the variable were varied over its full range. Note the y-axis scales differ depending on prevalence of the taxa. 15

List of Tables

Table 1: Species list and relative abundance of species at the 2269 sites initially used from the NZFFDB. 5

Table 2: Proportion of sites where each of the modelled species occurred and the critical threshold and area under curve (AUC) results for ROC analysis from N fold crossvalidation. 10

Table 3: Ranking of importance of the 35 predictor variables from the River Environment Classification and Freshwater Environments of New Zealand databases used for predicting fish communities. 12

Table 4: The 35 predictor variables from the River Environment Classification (REC) and Freshwater Environments of New Zealand (FWENZ) databases used for predicting fish communities grouped by upstream, segment and downstream influence. 17

Executive Summary

- A predictive model was developed for the wadeable streams ($\leq 4^{th}$ order) of the Waikato Region using existing fish presence/absence data.

- The environmental variables used to generate the predictions came from two existing datasets that were collated in a GIS format from other existing databases.

- The predictive model was extensively evaluated and iteratively optimised to maximise predictive accuracy.

- The levels of accuracy for the model were good to outstanding and exceeded those from comparable North Island regional models.

- The predictions from the model were then expanded out over the entire regional stream network to give a predictive map of fish assemblages.

- These predicted fish assemblages were then used to create a predictive map of IBI scores for the wadeable streams over the entire region.

- Notwithstanding the good evaluation results the model could be improved with more data, both more fish surveys and environmental data with higher resolution.

- The current predictions can be used to develop predictive maps of the distribution of 13 freshwater fish species for mapped wadeable streams throughout the Waikato Region.

1 Introduction

To model freshwater fish spatial occurrence in the Waikato Region data were extracted from the New Zealand Freshwater Fish Database (NZFFDB; McDowall and Richardson, 1983) and supplemented with data from recent surveys. The fish assemblage data for each site were associated with their corresponding catchment level geospatial landuse, geomorphologic, and climatic data in a geographic information system (GIS) to predict fish occurrence in the region in wadeable streams (defined as less than or equal to 4th order as identified by the REC). To predict the occurrence of each species at a site from a common set of predictor variables we used an artificial neural network, to produce a single model that predicts the entire fish assemblage at a site in one procedure. After extensive development and evaluation, the predictive model was then extended to fill in the gaps between the surveyed sites using a GIS river network to give a spatial map of species probability of occurrence for the wadeable streams over the entire region. The predictive map tool described here has been named Point-Click-Fish and has been developed for a number of New Zealand regions. The predicted fish assemblages were then run through the Waikato Index of Biotic integrity (IBI) (Joy & Death 2004; Joy, 2005) to calculate the IBI score for each wadeable segment of the river network.

The predictive maps of freshwater fish species distribution have been identified as an important tool for freshwater resource management in New Zealand (McLea & Joy, 2004). The maps allow non-specialist staff and others to instantly get information on the expected fish fauna for any site on any stream in the region. This spatial occurrence data can also be linked to extra information such as conservation status and habitat requirements, such as the climbing ability of species with a high probability of being present (McLea & Joy, 2004). The predictive IBI model has the potential for 'scenario modelling' where different management options can be modelled and changes in biotic integrity mapped out over the regional stream network.

1.1 Background

This capability to accurately predict where fish are likely to occur on a regional basis has many potential uses within regional councils for bioassessment, biodiversity assessment, resource management and conservation. The recent availability of a large database of geological, climatic and landuse information coupled with recent developments in computing power has made it possible to produce accurate maps of fish occurrence (Joy & Death, 2004). The combination of large amounts of GIS landscape data and fish distribution records mean that New Zealand has an almost unique opportunity to pioneer this process as the cost of this GIS data would rule out this taking place overseas.

1.2 Using artificial neural networks for spatial modelling

Artificial neural networks have 2 main advantages over almost any other modelling approach currently available 1) they can model both linear and non-linear relationships and 2) they can model the entire assemblage in one model. The neural network uses an iterative learning approach where, after each iteration (epoch), the difference between the predicted results and the actual results are compared and the mathematical relationship is updated each time based on minimising error (getting the answer right). Thus, the learning process is similar to human learning. For example when learning addition or multiplication, if you were not given the rules first, you would learn by looking for patterns. Each time an example is given the pattern between the number being added and the answer is observed until a definite pattern is learned and can be applied to further examples (providing enough examples are given). In the same way the neural network learns the relationship between the habitat variables and the fish presence/absence and develops an algorithm to classify new sites. When the

network is presented with a new example (where the answer is unknown) it is able to make a prediction using the rules it has learnt from the previous examples. Thus, in summary the patterns are learned from the fish/environment dataset given to 'train' the model. After fine-tuning, the model is then presented with the data for the entire regional wadeable stream network and the predictions are made then mapped out over the region.

This report outlines the development of a predictive model using data for 1967 wadeable stream sites in The Waikato region. The associations between fish assemblages at these sites and physicochemical data related to these sites (extracted from a number of existing databases) were modelled using a neural network. The statistical model was fine-tuned and then interrogated to give the variables a ranking based on their importance in predicting the assemblages and how the predictions related to individual variables. Next, the predictions were mapped out over the entire stream network to give a predictive map of fish assemblages. An index of biotic integrity was calculated for each wadeable segment and these were also mapped out over the entire region (see Fig 2 and data supplied as GIS layer).

2 Methods

2.1 Predictive model construction

A general outline of the development of the predictive model is given next and is shown diagrammatically in Fig. 1 (for technical details on model construction see Appendix 1). The presence/absence data for the 13 species present at more than 3% of sites were used to develop the predictive model. The 3% threshold was used as species present at only a few sites are very difficult to model because there a so few examples of their habitat requirements.

2.2 Data sources

2.2.1 Fish

The data on freshwater fish presence and absence came from the New Zealand Freshwater Fish Database (NZFFDB) (McDowall & Richardson, 1983). The NZFFDB sites selected were those including electrofishing, trapping and spotlight surveys giving 2269 sites. Where species identification was not definite (e.g. indeterminate bully species) these site records were removed as were sites where stream order was greater than order 4; this reduced the database to 1967 sites. Thirteen taxa were recorded at more than 3 % of the 1967 sites (taxa occurring at less than 3 % of sites were not included) (Table 1; Fig. 3).

Two non-migratory species are recorded in the Waikato Region; Cran's and upland bullies, but only 4 records were for upland bullies so these were discarded. Exotic species other than salmonids were not included in modelling because their distribution is generally related to areas where they have been released rather than to habitat requirements. The survey sites gave a relatively representative coverage of the region except for the southern, and the very northernmost part of the region (Fig. 4).

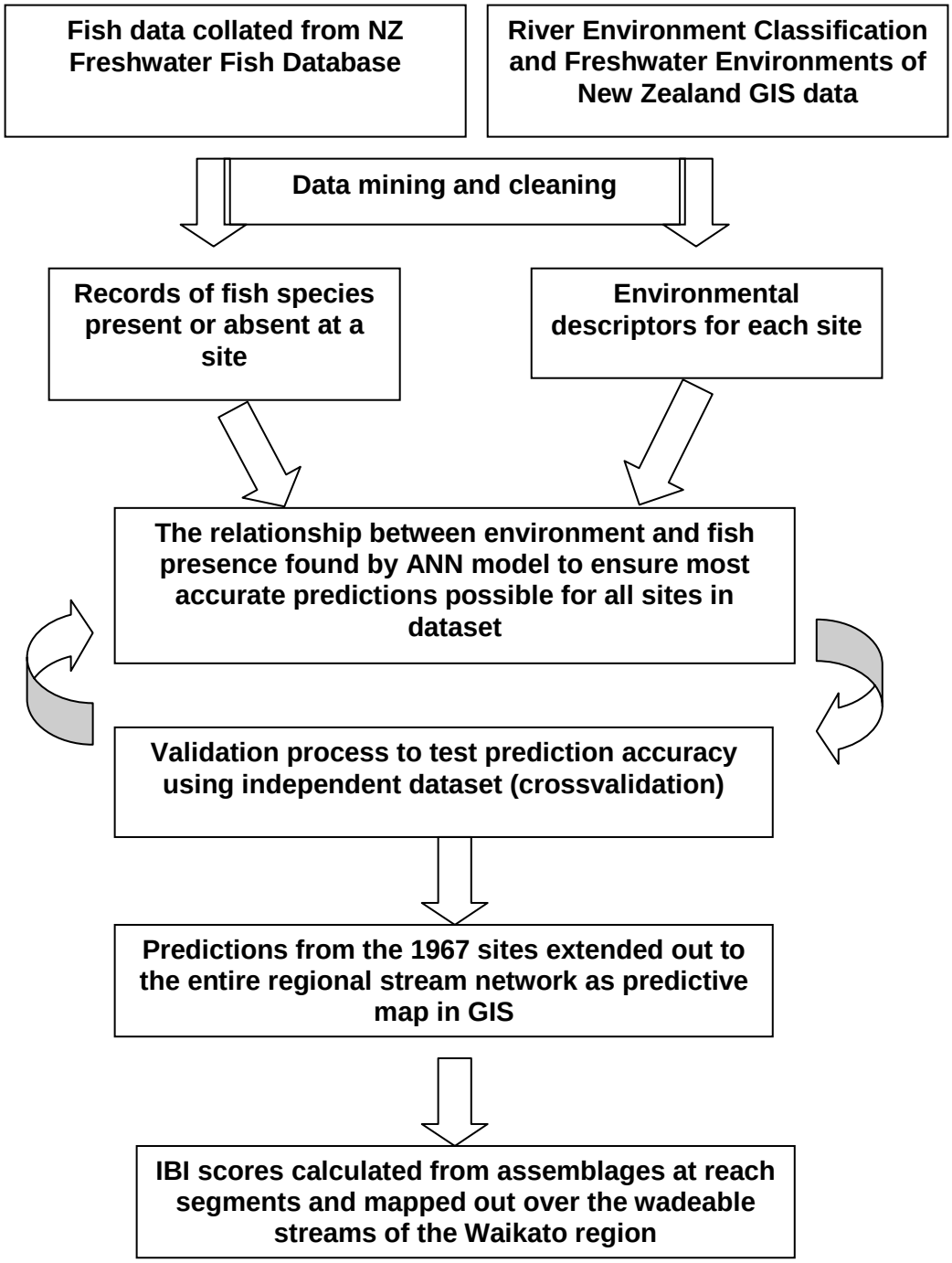

Figure 1: Flow chart for predictive IBI development.

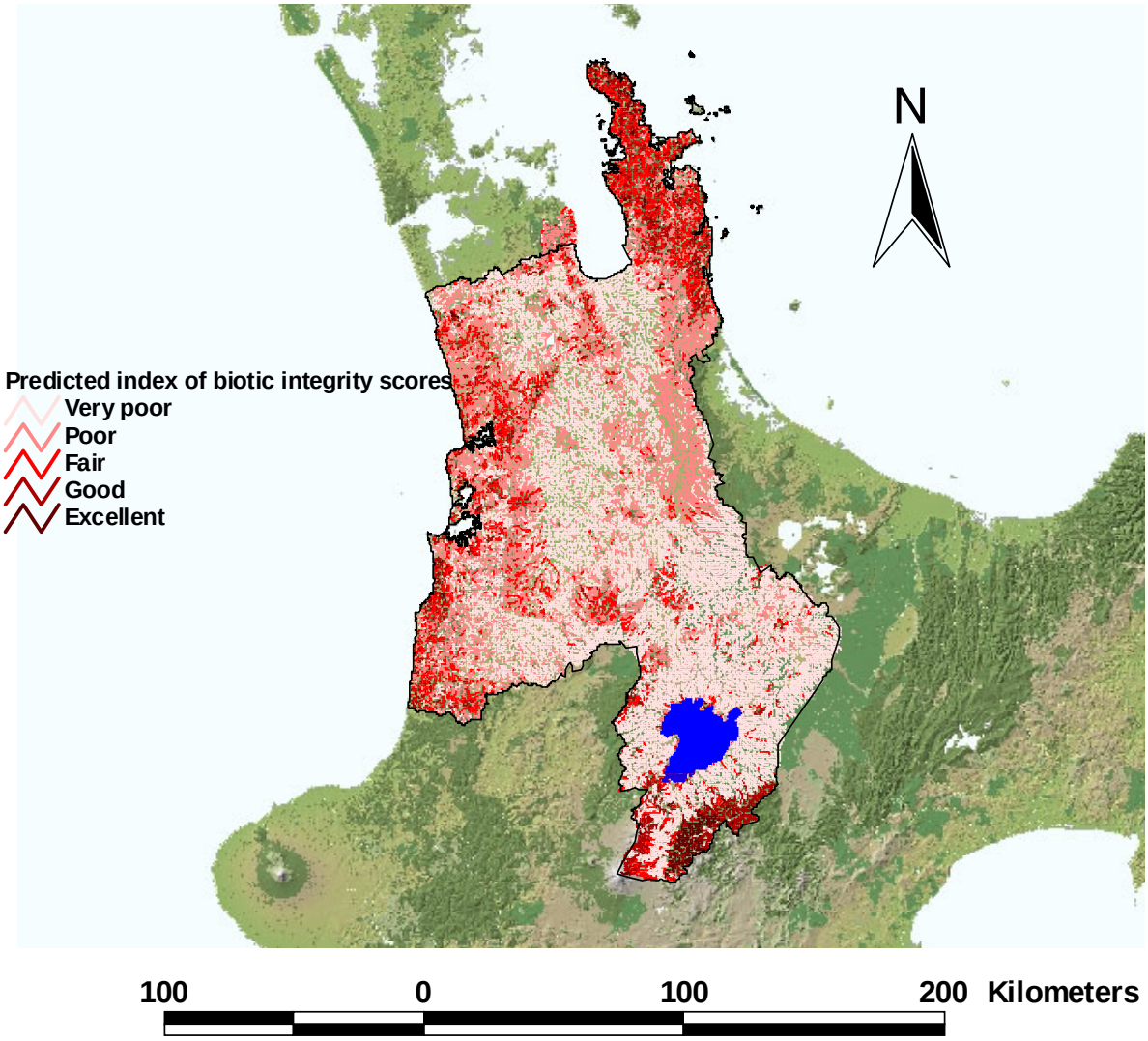

Figure 2: **Map showing the predicted IBI scores over the Waikato Region for streams of 4th order and less.**

Table 1: Species list and relative abundance of species at the 2269 sites initially used from the NZFFDB. After sites with unidentified species were removed, 1967 sites were available for model building. Twenty-seven sites were recorded as nil with no fish recorded and these were also removed. Species marked with an asterisk were not used in modelling.

Scientific name	Common name	Number of sites	Percentage of sites
Anguilla dieffenbachii	Longfin eel	1117	49.25
Anguilla australis	Shortfin eel	719	31.70
Gobiomorphus cotidianus	Common bully	540	23.81
Gobiomorphus huttoni	Redfin bully	458	20.19
Galaxias fasciatus	Banded kokopu	425	18.74
Galaxias maculatus	Inanga	319	14.07
Cheimarrichthys fosteri	Torrentfish	290	12.79
Gobiomorphus basalis	Cran's bully	282	12.43
*Anguilla spp.**	Unidentified eel	274	12.08
Retropinna retropinna	Common smelt	263	11.60
Oncorhynchus mykiss	Rainbow trout	257	11.33
Salmo trutta	Brown trout	160	7.05
*Gambusia affinis**	Gambusia	103	4.54
*Carassius auratus**	Goldfish	102	4.50
Galaxias brevipinnis	Koaro	85	3.75
*Ameiurus nebulosus**	Catfish	81	3.57
Galaxias argenteus	Giant kokopu	72	3.17
*Gobiomorphus spp.**	Unidentified bully	69	3.04
*Mugil cephalus**	Grey mullet	52	2.29
*Cyprinus carpio**	Koi carp	50	2.20
*Geotria australis**	Lamprey	47	2.07
*Galaxias spp.**	Unidentified galaxiid	41	1.81
*Scardinius erythrophthalmus**	Rudd	41	1.81
*Neochanna diversus**	Black mudfish	32	1.41
*Galaxias postvectis**	Shortjaw kokopu	23	1.01
*Gobiomorphus gobioides**	Giant bully	22	0.97
*Salmo**	Unidentified salmonid	16	0.71
*Gobiomorphus hubbsi**	Bluegill bully	15	0.66
*Aldrichetta forsteri**	Yelloweye mullet	9	0.40
*Ctenopharyngodon idella**	Grass carp	9	0.40
*Gobiomorphus breviceps**	Upland bully	4	0.18
*Mugil**	Unidentified mullet	4	0.18
*Poecilia reticulata**	Guppy	4	0.18
*Rhomboslea retiaria**	Black flounder	4	0.18
*Salvelinus fontinalis**	Brook char	4	0.18
*Grahamina sp.**	Estuarine triple fin	3	0.13
*Galaxias divergens**	Dwarf galaxias	1	0.04
Perca fluviatilis*	Perch	1	0.04

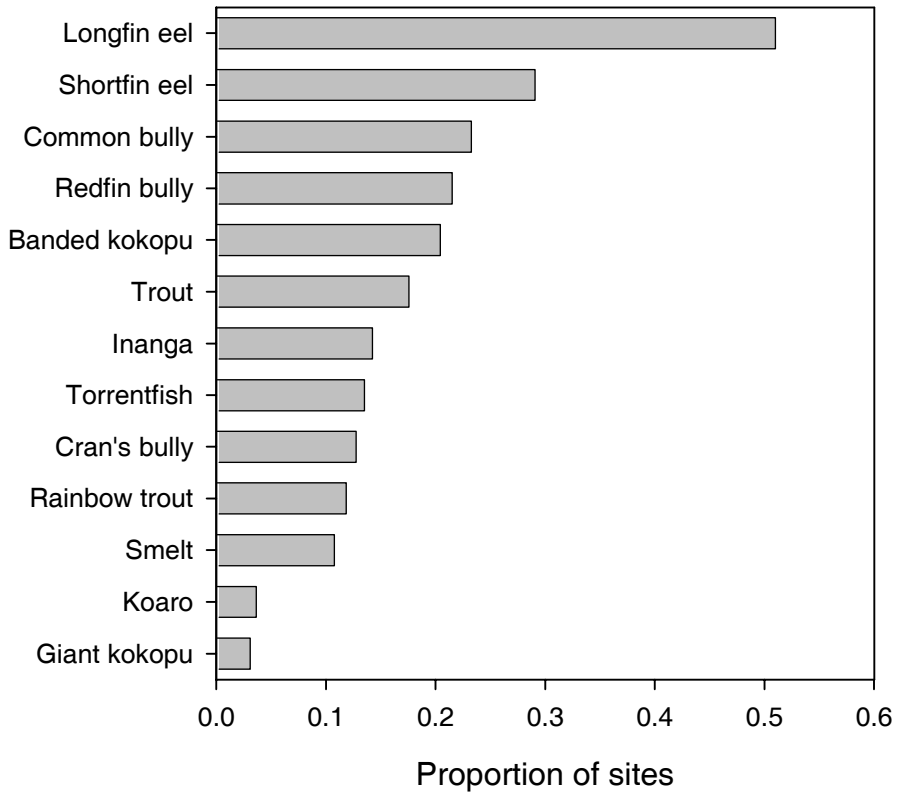

Figure 3: Proportion of the 1967 sites in the Waikato region at which each of the 13 modelled species were found.

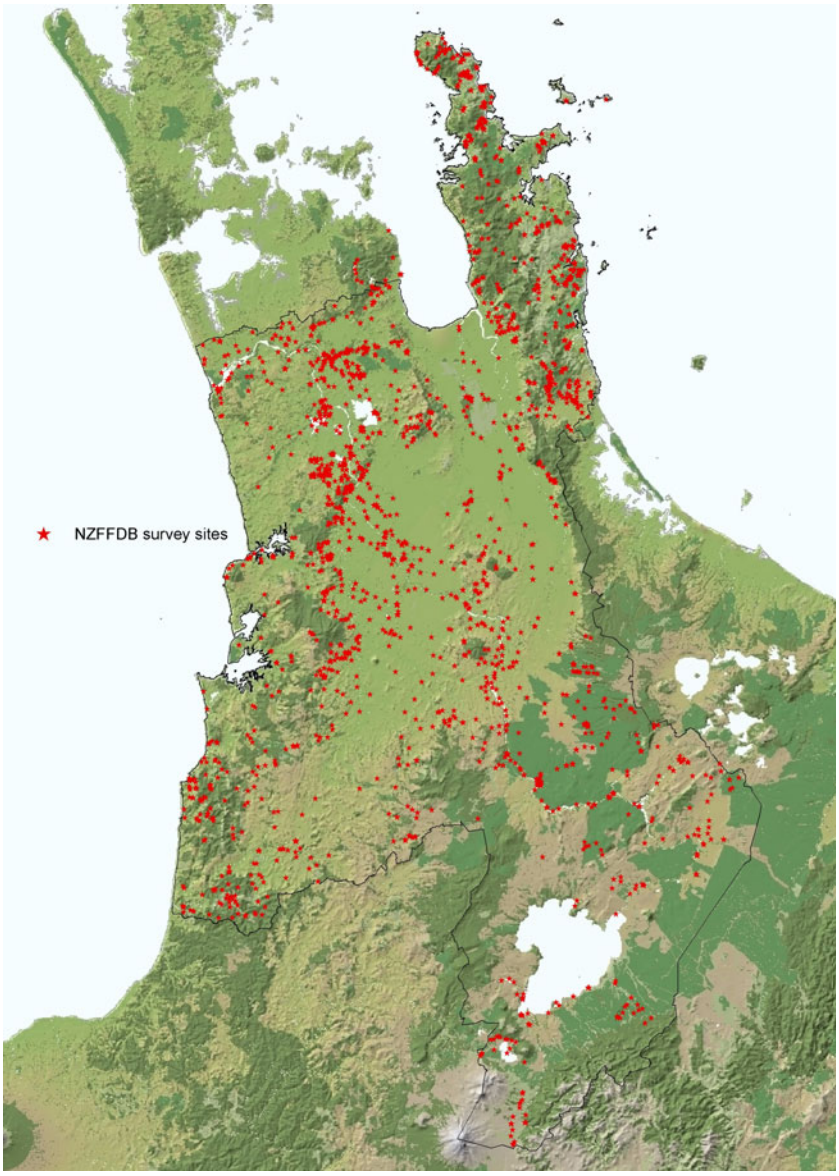

Figure 4: Site map showing the Waikato sample sites from the New Zealand Freshwater Fish Database.

2.2.2 Habitat

The habitat data for the predictions came from two sources; the River Environment Classification (REC; Snelder *et al.*, 1998) and the Freshwater Environments of New Zealand (FWENZ; Wild *et al.*, 2005) (see appendix for details of FWENZ variables used). The REC data used was the 'raw data'; that is, the proportion of the catchment covered by each variable rather than the categorical classification classes. The FWENZ data was developed for a separate classification process initiated by the Department of Conservation to classify rivers of national importance in parallel with the Land Environments of New Zealand (LENZ) classification process. Significantly, the FWENZ data was developed specifically for biological classification as opposed to the more physical/stream morphology emphasis in the REC. First the data were collated from the databases (see below for database details) then checked, screened, processed and formatted. In this process any of the environmental variables with a Pearson correlation co-efficient greater than 0.8 were removed from the list of potential variables. All reaches listed as stream order 5 or greater were removed as these were considered to be non-wadeable and so were not included in the data extraction process.

2.3 Model architecture and number of variables

An iterative approach was taken to deciding on the number of variables used and the model architecture. A combination of the number of hidden nodes and epochs was trialed using the average Cohen's kappa score as the criterion for deciding on the best combination (see model validation section below for more details). The entire dataset of environmental variables (minus highly correlated variables) was used initially and when a good architecture was found after iteration the best 35 variables were identified using the quantification process described below. The iterative approach above was again employed with the reduced variable set and the final architecture was determined.

2.4 Validation with independent data

Independent data were used to ensure that the model was not overtrained (this is where the model learns the training data so well that it is not useful on other data) using *N*-fold crossvalidation (also known as leave-one-out validation). This process involves taking out the first site in the list, building the model with the rest of the sites, and then testing the held out site using this model. The site is then put back and the next site taken out, again the model is built and the held out site is run through the model, and so on until all sites are tested and the results presented are for these held out sites. This effectively means that the model is tested on an independent dataset equivalent in size to the original dataset.

2.5 Model evaluation

Assessing the accuracy of distribution models is difficult because of the relative rareness or commonness of individual species. For example, if a fish is present at only 5% of the sites and a model is produced that predicts it will never occur, then that model will be assessed as being correct 95% of the time (or have a classification success of 95%). Obviously this result is misleading, and the results must be considered cautiously. One good way to judge the accuracy of predictive presence/absence models is to use a measure called Cohen's kappa (Fielding & Bell, 1997). This measure is independent of species prevalence and can be thought of as the percentage improvement in discrimination over chance.

A further difficulty with evaluating predictive presence/absence models is the probability cut-off used, because the model predictions are continuous values between 0 and 1 rather than either presence or absence. Traditionally the view has been that a predictive probability > 0.5 is considered to be present and a prediction less than 0.5 as absent, however, a better way to decide on the best cut-off to is to use a Receiver Operator Plot (ROC) (Zweig & Campbell, 1993). This process creates a curve based on all probability thresholds to find the best one to use to maximise prediction success. Furthermore, the area under the ROC can be used to estimate the accuracy of the predictions independent of the threshold used. The area under curve (AUC) values have been quantified by Hosmer & Lemeshow, (2000) who found that an AUC > 0.9 is outstanding discrimination; and 0.8 - 0.9 is excellent; an AUC 0.7 - 0.8 is acceptable discrimination, an AUC less than 0.6 is poor discrimination, and 0.5 and or represents no discrimination.

To optimise the predictive accuracy of the model a number of combinations of neural network architecture and the number of iterations were tried until the best output was achieved. This architecture was then used as the final model to predict fish communities over the entire regional stream network.

2.6 Quantifying predictor variable contributions

To determine the relative importance of each predictor variable we used the connection weights method (Olden and Jackson 2002). To calculate connection weights, the

product of the input-hidden and hidden-output connection weights between each input neuron and output neuron were summed across all hidden neurons using the raw connection weights (Olden and Jackson 2002). The ANN model was run 100 times and the average relative contribution of each variable was recorded from each run to rank the importance of the variables.

2.7 Sensitivity analysis

Sensitivity analysis was used to visualize the relationship between each of the species and the environmental variables. This is achieved by holding all the other variables at their mean values and then varying the variable of interest throughout its full range and plotting the relationship (Ozesmi & Ozesmi, 1999). The probability of encountering the individual fish species of interest can be calculated and plotted over the entire range of any variable of interest.

2.8 Predictive IBI

Of the 50697 stream reach segments identified to the Waikato Region, 47024 were less than or equal to 4th order (as identified by the REC) and the predicted fish assemblages for these segments were passed to the IBI model and the scores calculated. The IBI scores for the regional stream network were then mapped out.

3 Results

3.1 Network architecture and variables

The final architecture used for the predictive model contained 35 hidden nodes and was run for 500 epochs. Thirty five predictor variables were used after the ranking and quantification process (see Table 3 for variable list).

3.2 Model evaluation

3.2.1 Species comparison

The final predictive model after evaluation had a high level of accuracy; with an average crossvalidated correct classification rate of 85% (Table 2. Figure 5). The average Cohen's kappa score was 0.45 (range 0.30 – 0.67) which means that overall the predictions were at least 45% better than chance. These results are very conservative as they were obtained using *N*-fold crossvalidation, which effectively tests the predictions on another independent dataset equivalent in size to the original dataset. The evaluation measures given in Table 2 use the thresholds from the ROC analysis. The area under curve results show that 10 of the 13 taxa have prediction accuracies classed as outstanding or excellent (that is an AUC > 0.8; Hosmer & Lemeshow, 2000). The other 3 species have predictions classed as acceptable.

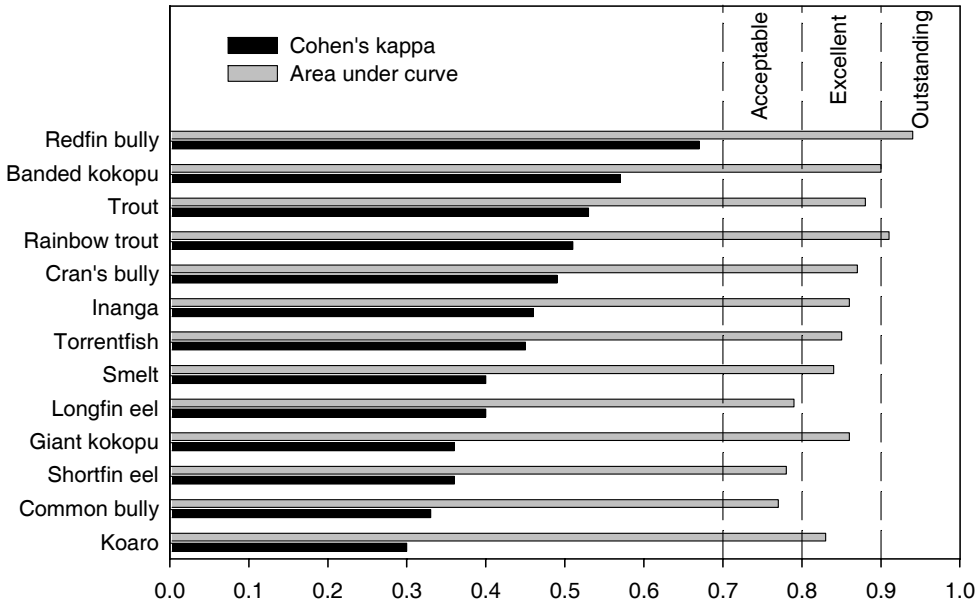

Figure 5: Crossvalidated Cohen's kappa results for the 13 species see text for details on classifications.

Table 2: Proportion of sites where each of the modelled species occurred and the critical threshold and area under curve (AUC) results for ROC analysis from N fold crossvalidation (see text for details). AUC results > 0.8 (bold) are considered excellent discrimination or > 0.9 (bold*) as outstanding discrimination (Hosmer & Lemeshow, 2000).

Species	Prevalence	Critical threshold	% correct	Cohen's kappa	Area under curve
Shortfin eel	0.29	0.43	0.75	0.36	0.78
Longfin eel	0.51	0.44	0.7	0.4	0.79
Torrentfish	0.14	0.30	0.86	0.45	**0.85**
Giant kokopu	0.03	0.54	0.96	0.36	**0.86**
Koaro	0.04	0.15	0.95	0.3	**0.83**
Banded kokopu	0.20	0.23	0.84	0.57	**0.90**
Inanga	0.14	0.40	0.86	0.46	**0.86**
Cran's bully	0.13	0.36	0.88	0.49	**0.87**
Common bully	0.23	0.30	0.73	0.33	0.77
Redfin bully	0.22	0.45	0.88	0.67	**0.94***
Rainbow trout	0.12	0.20	0.88	0.51	**0.91***
Smelt	0.11	0.38	0.89	0.4	**0.84**
Brown trout	0.18	0.34	0.86	0.53	**0.88**
Mean	0.18	0.35	0.85	0.45	**0.85**

3.2.2 Assemblage comparison

Another way to evaluate the match between predictions and reality is to compare how the predicted assemblages compare with the observed assemblages. To do this the simple matching coefficient was used, this is the percentage similarity between observed and expected assemblages. The simple matching coefficient results are shown in Figure 6. Eight hundred and sixty sites gave a perfect match (100% similarity), a further 454 sites were 90% or more similar and 523 sites had an 80% or better match. If those three groups are added together then more than 90% the sites have an 80 % or better match between observed and expected communities. Because this evaluation is based on what is effectively an independent dataset, we can expect that this level of accuracy would be similar for any new sites.

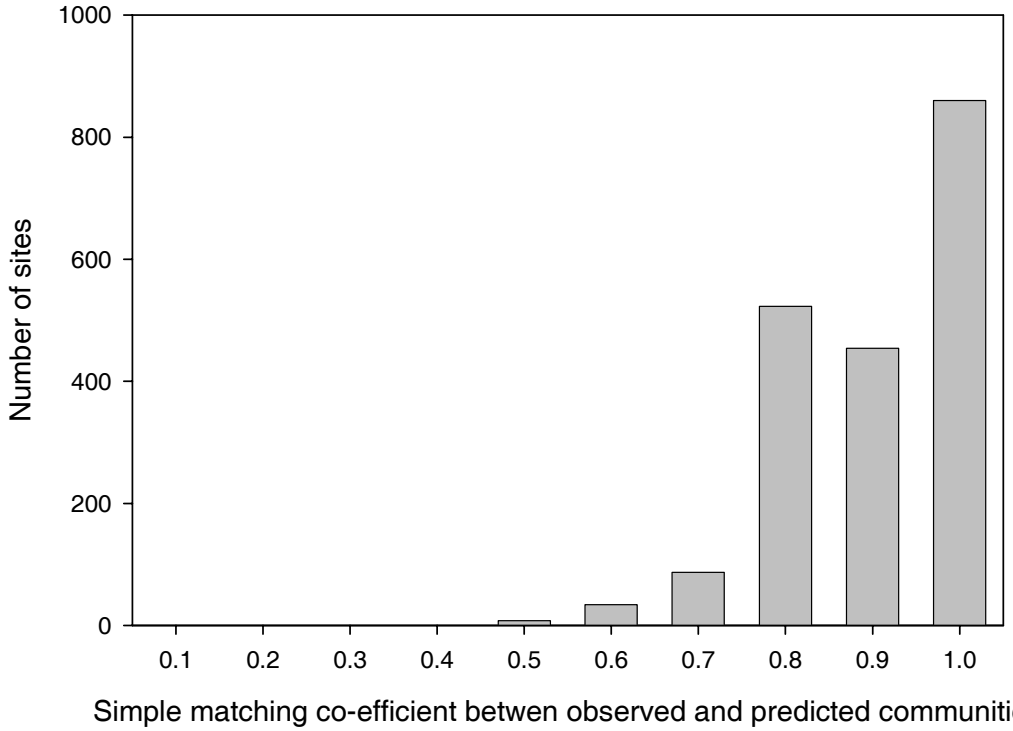

Figure 6: Histogram of crossvalidated simple matching coefficients for the match between observed and predicted fish assemblages using the ANN model at the 1967 Waikato sites. Matching coefficient values >0.8 indicate more than 80% matching between observed and expected communities.

3.3 Predictor variable importance

The relative influence of predictor variables was calculated from the connection weights in the neural network and then ranked (Table 3). The variable rankings were dominated by the variables calculated from upstream catchments and weighted by rainfall. Three of the first 4 variables are climate related - temperature and rainfall. The highest ranking variable (Usaveslope) was calculated from the average slope of the stream above the site from a 30m digital elevation model (DEM) and was weighted by the rainfall in the catchment area. Stream order was the next ranked variable followed by phosphorus. The first landcover variable (upstream indigenous forest) came in at 7 in the ranking followed by the first of the segment variables - the slope of the segment where the site is located. Moving down the list the variables next in the ranking are a mixture of geology and climate until the first of the downstream variables (the distance from the site to the sea). The next variables are segment shade, upstream pasture, raindays and the presence of a downstream floodgate. The next downstream variable in the ranking is the presence of a dam. The subsequent variables are a mixture of upstream and segment landcover, other climate variables, and the presence of a downstream culvert.

Table 3: Ranking of importance of the 35 predictor variables from the River Environment Classification and Freshwater Environments of New Zealand databases used for predicting fish communities.

Rank	Abbreviation	Description of variable
1	Usaveslope	Runoff weighted catchment average slope calculated for 30m DEM grid
2	Usavtcold_q	Runoff weighted mean minimum July air temperature
3	Usraindays20_q	Runoff weighted catchment rain days (greater than 20mm/month)
4	Usraindays25_q	Catchment rain days (greater than 25mm/month)
5	Order	Stream order
6	Usphos_q	Runoff weighted catchment average of phosphorus
7	Usindigforest_q	% of catchment in LCDB category (indigenous forest)
8	Segmaxslope_grid	Maximum segment slope based on 30m DEM grid
9	Ushard_q	Runoff weighted catchment average of hardness (induration)
10	Ussolarradsum	Runoff weighted December catchment solar radiation
11	Segequitwin	Current wintertime equilibrium temperature
12	Topvolcsof	Proportion of catchment volcanic soft
13	Topvolchar	Proportion of catchment volcanic hard
14	Topgrey	Proportion of catchment in greywacke
15	Distsea	Distance from the coast
16	Segshade	Estimate of current segment shade
17	usPastoral_Q	% of annual runoff from LCDB category (pastoral)
18	Segpastoral	% of riparian area in LCDB category (pastoral)
19	Usraindays200_q	Runoff weighted catchment rain days (greater than 200mm/month)
20	Fld_gat	Flood gate downstream
21	Usexoticforest	% of annual runoff from LCDB category (exotic forest)
22	Segindigforest	% of riparian area in LCDB category (indigenous forest)
23	Segslope	Average segment slope
24	Dam	Presence of a dam downstream
25	Lcdbarea	Catchment area from LCDB
26	Segexoticforest	% of riparian area in LCDB category (exotic forest)
27	Usscrub_q	% of annual runoff from LCDB category (scrub)
28	Aveelev	Average segment elevation
29	Segscrub	% of riparian area in LCDB category (scrub)
30	Segbare	% of riparian area in LCDB category (bare)
31	Cvt	Presence of a culvert downstream
32	Reachlen	Segment length
33	Usavtwarm_q	Runoff weighted Mean January air temperature
34	Ussteep_q	% annual runoff volume from area of catchment with slope > 30°
35	Uslowgrad_q	% annual runoff volume from area of catchment with slope < 30°

3.4 Sensitivity analysis

The direction of influence of any of these variables on individual species can be calculated using sensitivity analysis, which is obtained by holding all the other variables at their mean values, and then varying the variable of interest throughout its full range and plotting the relationship. With thirty-five variables and 13 species there are a large number of combinations but a few examples for 12 species are given in Figures 7 and 8. The first set of plots (Fig. 7) show the influence of average catchment phosphorus weighted by rainfall on the probability of capture for the 12 species. The plots show that the probability of occurrence reduces with increasing catchment area in phosphorus for most of the species except Cran's and common bullies, rainbow trout and smelt whose probability of occurrence peaks at moderate phosphorus levels. The next set of plots (Fig. 8) show the relationship between the probability of occurrence and the proportion of the segment in pastoral farming. The plots show the increasing likelihood of shortfin eels, torrentfish, inanga, Cran's and common bullies occurring with increasing proportions of pastoral landuse. The opposite is revealed for longfin eels, giant kokopu, koaro, banded kokopu, redfin bullies and smelt with decreasing probability of occurrence with increasing phosphorus. When considering these plots it is important to note that these relationships are not necessarily causative as many of the variables are inter-correlated with landuse patterns.

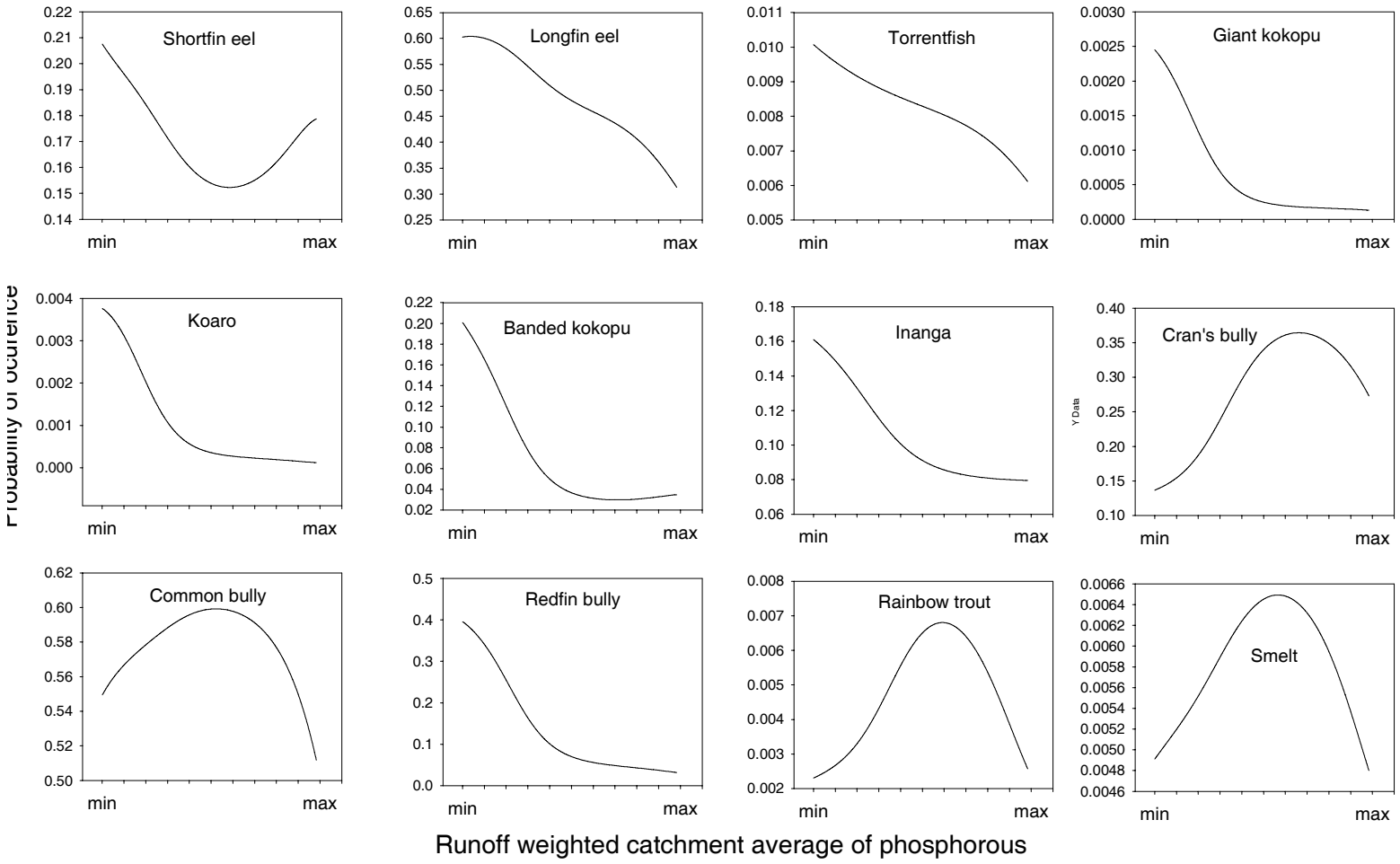

Figure 7: Sensitivity analysis for catchment phosphorus versus probability of occurrence. In each plot (one for each taxon) all other variables were set to their mean values and the catchment proportions of the variable were varied over its full range. Note the y-axis scales differ depending on prevalence of the taxa.

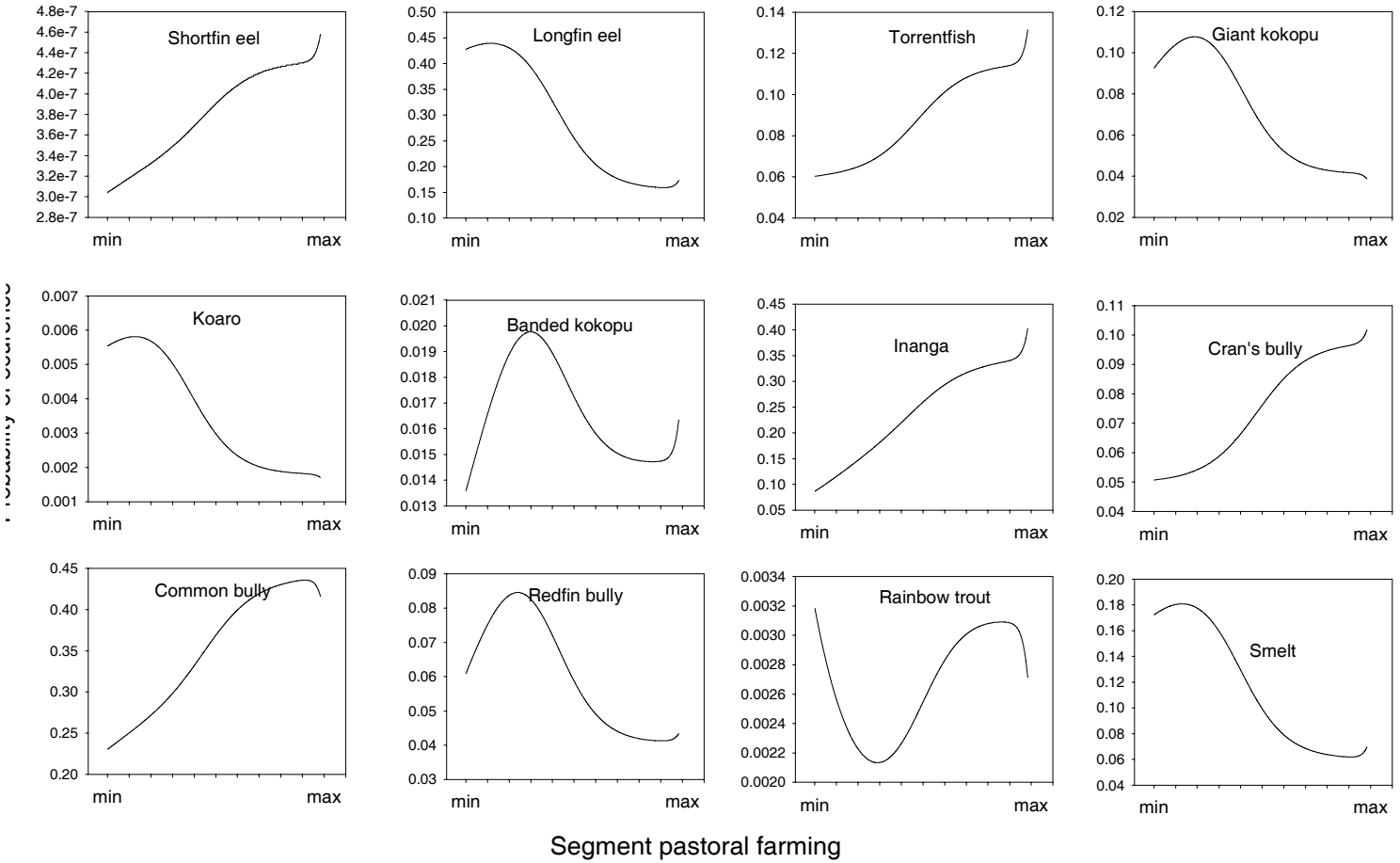

Figure 8: Sensitivity analysis for the proportion of the stream segment in pastoral farming versus probability of occurrence. In each plot (one for each taxon) all other variables were set to their mean values and the catchment proportions of the variable were varied over its full range. Note the y-axis scales differ depending on prevalence of the taxa.

4 Discussion

Regional models predicting fish assemblages over the entire stream network have been developed for three North Island regions; Wellington, Hawkes Bay and now for wadeable streams in Waikato. Similar although lower levels of prediction accuracy have been found for the other models but the prediction accuracies are higher for the regional models than for a similar all North Island model (MKJ unpublished data). This suggests that there are regional differences in the relationships between variables and assemblages among North Island regions or that the relative importance of variables varies between the different regions.

The high levels of predictive accuracy revealed in this model reveal the strong relationship between catchment scale variables and fish occurrence. In contrast, the predictions of stream invertebrate communities at the catchment scale were not as successful (MKJ unpublished data), suggesting that local or proximal scale variables are more important for the invertebrates. The strong influence of flow, slope, rainfall, temperature, stream-size and catchment vegetation on fish assemblages was revealed by the predictive model, as downstream or proximal variables were well down the variable ranking list.

The 1967 sites available for model building were considerably higher than for the Hawkes Bay Region where more than 600 sites were available and the Wellington Region where 379 sites were used. Obviously the more sites available the better the model will be, and the evaluation methodology used showed that the model performed very well. The most important factor is that the sites used cover all the possible stream habitat types that exist in the region. Also, the sampling methods used for most of the sites in the NZFFDB are not able to accurately represent stream fish assemblages in large or non-wadeable rivers, thus wadeable streams ($\leq 4^{th}$ order) only were used in this model.

4.1 Assemblage-environment relationships

The development of the FWENZ variables used as predictors here followed on from the Land Environments of New Zealand (LENZ) process. Discussion of their usefulness is beyond the scope of this report but Wild *et al.*, (2005) discuss these variables and their calculation in detail. The environmental variables associated with the fish assemblages were ranked by their importance but it is essential to note that this is importance when predicting whole assemblages and some variables will be more or less important when species are considered individually. Many of the environmental variables available to use as predictors are highly correlated with each other, for example landuse tends to change with elevation and distance from the coast, as do rainfall and temperature. The removal of highly correlated variables during data-screening lessened the influence of these correlations. Many of the FWENZ variables were weighted by flow by using average rainfall multiplied by the area of the catchment. These flow weighted variables were generally the variables mostly selected by the model as the best predictors. The separation of variables into upstream, downstream and segment classes (Table 4) reveals the importance of upstream variables for predicting the stream communities. The first of the segment variables comes in at 8 in the ranking and the first of the downstream variables is not until 15.

Table 4: The 35 predictor variables from the River Environment Classification (REC) and Freshwater Environments of New Zealand (FWENZ) databases used for predicting fish communities grouped by upstream, segment and downstream influence.

Overall rank	Upstream variables
1	Runoff weighted catchment average slope calculated for 30m DEM grid
2	Runoff weighted mean minimum July air temperature
3	Runoff weighted catchment rain days (greater than 20mm/month)
6	Runoff weighted catchment average of phosphorus
7	% of catchment in LCDB category (indigenous forest)
9	Runoff weighted catchment average of hardness (induration)
10	Runoff weighted December catchment solar radiation
12	Proportion of catchment volcanic soft rock
13	Proportion of catchment volcanic hard rock
14	Proportion of catchment in greywacke
17	% of annual runoff from LCDB category (pastoral)
19	Runoff weighted catchment rain days (greater than 200mm/month)
21	% of annual runoff from LCDB category (exotic forest)
25	Catchment area from LCDB
27	% of annual runoff from LCDB category (scrub)
33	Runoff weighted mean January air temperature
34	% annual runoff volume from area of catchment with slope > 30°
35	% annual runoff volume from area of catchment with slope < 30°
Segment variables	
5	Stream order
8	Maximum segment slope based on 30m DEM grid
11	Current wintertime equilibrium temperature
16	Estimate of current segment shade
18	% of riparian area in LCDB category (pastoral)
22	% of riparian area in LCDB category (indigenous forest)
23	Average segment slope
26	% of riparian area in LCDB category (exotic forest)
28	Average segment elevation
29	% of riparian area in LCDB category (scrub)
30	% of riparian area in LCDB category (bare)
32	Segment length
Downstream variables	
15	Distance from the coast
20	Flood gate downstream
24	Presence of a dam downstream
31	Presence of a culvert downstream

4.2 Species-environment relationships

The individual species-environment relations are summarised by the sensitivity analysis. Because of the 13 species and 35 variables there are a large number (455) of associations to analyse so this must be restricted to particular combinations of interest (Figs. 7 & 8).

4.3 Limitations of the predictive model

The predictions of the model are limited by the range of sites used to build the model and the nature/accuracy of the predictor variables. The assumptions behind a regional predictive model are:

1. that all main habitat types existing in the region are found in the sites used in training the model and;
2. that all the important variables influencing the assemblages being modelled are available as predictor variables and that the variables provided accurately represent environmental conditions prevailing in the catchments.

The plot showing the placement of survey sites (Fig. 4) shows a reasonably good spatial coverage which suggests that most of the available habitats have been covered. The exceptions would be the very large and very small streams. Rivers larger than fourth order are problematic as they are rarely sampled and samples tend to only be representative of the edge of the river rather than the whole river at the site. Furthermore, the smaller streams are not represented by the REC drainage network. The current predictions can be used to develop predictive maps of the distribution of 13 freshwater fish species for mapped wadeable streams throughout the Waikato Region

4.4 Future data requirements

The model could be significantly improved with more accurate data. One of the most important predictor variables is landuse but there is at present no data available discriminating farming intensity with sufficient detail. The data from the LCDB has just one pastoral farming class covering all farm types although the recently available LCDB2 has two pastoral classes which may be an improvement. Other regional finer scale data such as that available using Light Detection and Ranging (LIDAR) are likely to be very useful for refining the accuracy of predictor variables. Work is currently underway at Environment Waikato to develop more detailed pastoral landuse predictor variables.

References

Bishop, C.M. 1995: *Neural networks for pattern recognition.* Oxford University Press, New York

Fielding, A.H. and Bell, J.F. 1997: A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation.* 24:38-49.

Foody, G. M. 1992: On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing.* 58:1459-1460.

Hosmer, D.W. and Lemeshow, S. 2000: *Applied logistic regression*, 2nd ed. Wiley-Interscience, New York

Joy, M.K. and Death, R.G. 2004: Predictive modelling and spatial mapping of freshwater fish and decapod assemblages: an integrated GIS and neural network approach. *Freshwater Biology.* 49:1036-1052.

Joy, M.K. 2005: *A Fish Index of Biotic Integrity (IBI) for the Waikato Region.* Massey University, Palmerston North

Krebs, C.J. 1999: *Ecological methodology.* 2nd ed. Benjamin/Cummings, Menlo Park, Calif.

Kurkova, V. 1992: Kolmogrov's theorem and multilayer neural networks. *Neural Networks.* 5:501-506.

Lek, S.; Belaud, A; Dimopulous, I; Lauga, J and Moreau, J. 1996: Improved estimation, using neural networks, of the food consumption of fish populations. *Marine and Freshwater Research* 46:1229-1236.

Manel, S.: Williams, H.C. and Ormerod, S.J. 2001: Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology.* 38: 921-931.

McDowall, R.M. and Richardson, J. 1983: *The New Zealand freshwater fish survey-a guide to input and output*. Report No. 12. Ministry of Agriculture and Fisheries, Wellington

McLea, M. and Joy, M.K. 2004: Point-Click-Fish A management tool for freshwater fish. *Paper presented at the 7th International River Symposium, Brisbane, Australia, 1-3 September 2004.*

Oberdorff, T.; Pont, D.; Hugueny, B. and Chessel, D. 2001: A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. *Freshwater Biology* 46:399-415.

Olden, J.D. and Jackson, D.A. 2000: Torturing data for the sake of generality: How valid are our regression models? *Ecoscience.* 7:501-510.

Olden, J.D. and Jackson, D.A. 2001: Fish-habitat relationships in lakes: Gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society.* 130:878-897.

Olden, J.D. and Jackson, D.A. 2002: A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology* 47:1976-1995.

Özesmi, S.L. and Özesmi, U. 1999: An artificial network approach to spatial habitat modelling with interspecific interaction*. Ecological Modelling.* 116:15-31.

Rumelhart, D.E.; Hinton, G.E. and Williams, R.J. 1986: Learning representations by back-propagating errors. *Nature.* 323:533-536.

Snelder, T.; Biggs, B.J.F.; Shankar, U.; McDowall, B.; Stephens, T. and Boothroyd, I.K.G. 1998: *Development of a system of physically based habitat classification for water resources management of New Zealand rivers.* National Institute of Water and Atmospheric Research (NIWA), Christchurch.

Titus, K.; Mosher, J.A. and Williams, B.K. 1984: Chance-corrected classification for use in discriminant analysis. *The American Midland Naturalist.* 111:1-7

Wild, M.; Snelder, T.; Leathwick, J.R.; Shankar, U. and Hurren, H. 2005: *Environmental variables for the Freshwater Environments of New Zealand River Classification,* Rep. No. CHC2004-086. National Institute of Water and Atmospheric Research (NIWA), Christchurch.

Zweig, M. and Campbell, G. 1993: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry.* 39:561-577

Appendix 1: Technical details on model construction validation and evaluation from Joy & Death (2004)

Artificial neural networks

Of the many neural network types available (see (Bishop, 1995) we chose a single hidden-layer feedforward multi layer perceptron trained using the backpropagation error algorithm (Rumelhart *et al.*, 1986). Specifically we used the resilient backpropagation 'Trainrp' (Matlab®) network training function that updates weight and bias values according to the resilient backpropagation algorithm (RPROP) during training. This network consisted of single predictor, hidden and output layers with a predictor neuron for each independent variable and an output neuron for each dependant variable. A single hidden layer was used because it reduces computation time and often produces similar results to networks with multiple hidden layers (Bishop, 1995; Kurkova, 1992). Each neuron in the network is connected to all neurons from adjacent layers. These connections among neurons are assigned a weight that dictates the intensity of the signal transmitted by the connection and the state of each input neuron is defined by the incoming signal transmitted by the input variable. The state of the other neurons is evaluated locally by calculating the weighted sum of the incoming values of the predictor variables of the previous layer. The network is trained with a back-propagation algorithm in which all weights are iteratively adjusted, with the goal of finding a set of connection weights that minimizes the error of the network. During the training process the observations are presented to the network sequentially and the weights adjusted depending on the magnitude and direction of the error. Learning rate and momentum vary as a function of the network error to ensure a high probability of global network convergence (see Bishop (1995) for details of the procedure described above). For further details of the mathematical aspects of ANNs and their use in ecological applications we refer readers to Lek *et al*. (1996) and Olden and Jackson (2001).

We determined the optimal number of hidden neurons and number of training epochs iteratively by comparing the performances of different networks. To achieve this we compared networks with 20 to 120 (in intervals of 20) hidden neurons and varied the number of epochs from 50 to 250 (in 50 epoch intervals) and then selected the combination that produced the greatest predictive accuracy based on the evaluation procedures outlined below.

We used crossvalidation (see model evaluation below for details) for this optimisation to ensure that the network was not 'overtrained'. Overtraining occurs when the network learns the training data extremely well but is not able to generalise well (Chatfield, 1995). After optimisation, the network consisted of 31 predictor neurons representing each of the independent predictor variables. The hidden layer consisted of 70 neurons and there were 18 output nodes, one for each of the dependant variables (the 18 species being modelled) and the training was run for 100 epochs. The independent variables were converted to *z* – scores prior to training.

Model evaluation

To evaluate the accuracy of the predictive model we used a leave-one-out cross-validation (jack-knife) method. This method involves excluding one observation, reconstructing the model and then predicting the response of the excluded observation. This process provides a nearly unbiased estimate of model accuracy (Olden, 2000). To assess the overall classification success of the model we first derived matrices of confusion (Fielding & Bell, 1997). A matrix of confusion tabulates the observed and predicted presence/absence patterns to provide a summary of the number of correct and incorrect classifications from the model. Using these matrices, five metrics of

prediction success were produced: (1) The overall classification accuracy of the model was measured as the percentage of sites where the model correctly predicted the presence/absence of each of the species. (2) The ability of the model to accurately predict species presence was assessed as model sensitivity (i.e. the percentage of site presences correctly predicted). (3) The ability of the model to correctly predict species absences was assessed and recorded as model specificity. (4) Cohen's Kappa coefficient of agreement (Titus *et al.*, 1984) was used to examine if model accuracy differed from expectation based on chance alone. We also calculated the Kappa *z*-score, 95 % confidence interval and *P*-value for each taxon using PROC FREQ (SAS 2000). Cohen's kappa is a robust model evaluation method that is relatively independent of species frequency of occurrence (Manel *et al.*, 2001); furthermore Cohen's kappa is a conservative estimate of prediction accuracy as the estimation of the agreement due to chance is underestimated (Foody 1992).

Rather than simply following the conventional decision threshold of 0.5 (e.g. Oberdorff *et al,* 2001) for deciding on species presence we constructed receiver-operating characteristic (ROC) plots to estimate the predictive ability of the model over all decision thresholds and find the optimal threshold (Fielding & Bell, 1997; Zweig & Campbell, 1993). The ROC plot is obtained by plotting the true positive proportion on the *y*-axis against the false positive proportion on the *x*-axis as the decision threshold is varied over the entire range between 0 and 1. The optimum threshold is chosen to maximise overall classification accuracy of the model assuming equal costs of misclassification of species presence/absence. The area under the ROC curve (AUC) is an index of accuracy as it provides a single evaluation measure independent of any particular threshold. Confidence intervals were obtained for the AUC from 999 bootstraps from observed and predicted values for each taxon. We used the AUC from the ROC plots as the fifth evaluation method and to find the optimal probability threshold for each taxon (Hosmer & Lemeshow, 2000; Zweig & Campbell, 1993). Finally, we compared the entire predicted communities (using crossvalidation) to those observed at the 1967 sites to measure the percentage of similarity between the observed and expected community using the simple matching coefficient, (Krebs, 1999).

Quantifying predictor variable contributions

To determine the relative importance of each predictor variable we used the connection weights method (Olden and Jackson 2002). To calculate connection weights the product of the input-hidden and hidden-output connection weights between each input neuron and output neuron were summed across all hidden neurons using the raw connection weights (Olden and Jackson 2002b). The ANN model was run 100 times and the average relative contribution of each variable was recorded from each run to rank the importance of the variables.

To elucidate the direction of relationship between predictions and predictor variables we used sensitivity analysis (Özesmi and Özesmi, 1999). This analysis involves varying one predictor variable over its entire range while holding all other variables at their mean value and examining the response of the output predictions. However, in this case where there are a large number of connections the process becomes rather cumbersome. Therefore, we selected some variables of interest (generally associated with fish distribution) that also had high contributions from the connections weights analysis as examples and examined the direction of influence for these variables. All the procedures described above were run on Matlab® version 6.5 using the Neural Networks toolbox.

Appendix 2: FWENZ variables used in model construction

Variable description	Variable Name	Units
Network from node	Nzfnode	dimensionless
NZREACH	NZREACH	dimensionless
Network to node	Nztnode	dimensionless
Segment maximum elevation based on 30m DEM	segMaxElev_Grid	m
Segment minimum elevation based on 30m DEM	segMinElev_Grid	m
X coordinate of catchment centroid	usXcentroid	m
Y coordinate of catchment centroid	usYcentroid	m
Segment length	segLen	m
Euclidean length of segment	segEucLen	m
Stream order	SegOrder	dimensionless
Segment average elevation based on 30m DEM	segAveElev_Grid	m
Elevation of upstream end of segment (From REC)	segUpElev	m
Elevation of downstream end of segment (From REC)	segDownElev	m
Total catchment area	usArea	m^2
Downstream Variables		
Average slope of downstream network	dsAveSlope	ratio
Distance to coast from segment	dsDistToSea	m
Maximum slope of downstream segments	dsMaxSlope	ratio
Maximum of maximum downstream grid slope	dsMaxSlope_Grid	angle - degrees
Segment Variables 1 (Climate)		
Average within segment mean minimum June air temperature	segAveTCold	$^oC*10$
Average within segment mean January air temperature	segAveTWarm	$^oC*10$
Current summertime equilibrium temperature	segEquiTSum	^oC
Historic summertime equilibrium temperature	segEquiTSum_Hist	^oC
Current wintertime equilibrium temperature	segEquiTwin	^oC
Historic wintertime equilibrium temperature	segEquiTwin_Hist	^oC
Segment December solar radiation	segSolarRadSum	W/m^2
Segment June solar radiation	segSolarRadWin	W/m^2
Table 4: Segment Variables 2 (Morphology)		
Maximum segment slope based on 30m DEM grid	segMaxSlope_Grid	angle
Segment sinuosity	segSinu	reachlen/euclen
Average segment slope	segSlope	ratio
Average within segment slope based on 30m DEM grid	segSlope_Grid	angle
Estimate of historic segment land cover	segVeg_Hist	Dimensionless
Estimate of current segment shade	segShade	Dimensionless
Estimate of historic segment shade	segShade_Hist	Dimensionless
Segment Variables 3 (Land cover)		
% of riparian area in LCDB category (wetland)	segWetland	%
% of riparian area in LCDB category (urban)	segUrban	%
% of riparian area in LCDB category (tussock)	segTussock	%
% of riparian area in LCDB category (miscellaneous)	segMiscLandCover	%
% of riparian area in LCDB category (pastoral)	segPastoral	%
% of riparian area in LCDB category (scrub)	segScrub	%
% of riparian area in LCDB category (bare)	segBare	%
% of riparian area in LCDB category (exotic forest)	segExoticForest	%
% of riparian area in LCDB category (indigenous forest)	segIndigForest	%

Variable description	Variable Name	Units
Upstream Variables 1 (Climate/Flow)		
Coefficient of variation of annual catchment rainfall	usAnRainVar	mm
Runoff weighted coefficient of variation of annual catchment rainfall	usAnRainVar_Q	mm
Runoff weighted catchment average slope calculated for 30m DEM grid	usAveSlope_Q	angle - degrees
Mean minimum July air temperature	usAvTCold	^{o}C*10
Runoff weighted mean minimum July air temperature	usAvTCold_Q	^{o}C*10
Mean January air temperature	usAvTWarm	^{o}C*10
Runoff weighted Mean January air temperature	usAvTWarm_Q	^{o}C*10
Catchment rain days (greater than 10mm/month)	usRainDays10	mean # days/yr
Runoff weighted catchment rain days (greater than 10mm/month)	usRainDays10_Q	mean # days/yr
Catchment rain days (greater than 100mm/month)	usRainDays100	mean # days/yr
Runoff weighted catchment rain days (greater than 100mm/month)	usRainDays100_Q	mean # days/yr
Catchment rain days (greater than 15mm/month)	usRainDays15	mean # days/yr
Runoff weighted catchment rain days (greater than 15mm/month)	usRainDays15_Q	mean # days/yr
Catchment rain days (greater than 20mm/month)	usRainDays20	mean # days/yr
Runoff weighted catchment rain days (greater than 20mm/month)	usRainDays20_Q	mean # days/yr
Catchment rain days (greater than 200mm/month)	usRainDays200	mean # days/yr
Runoff weighted catchment rain days (greater than 200mm/month)	usRainDays200_Q	mean # days/yr
Catchment rain days (greater than 25mm/month)	usRainDays25	mean # days/yr
Runoff weighted catchment rain days (greater than 25mm/month)	usRainDays25_Q	mean # days/yr
Catchment rain days (greater than 50mm/month)	usRainDays50	mean # days/yr
Runoff weighted catchment rain days (greater than 50mm/month)	usRainDays50_Q	mean # days/yr
Total annual runoff volume	usFlow	mm*m^2/yr
Mean annual low flow	usLowFlow	l/s
Annual potential evapotranspiration of catchment	usPET	mm
Runoff weighted annual potential evapotranspiration of catchment	usPET_Q	mm
December catchment solar radiation	usSolarRadSum	W/m^2
Runoff weighted catchment December solar radiation	usSolarRadSum_Q	W/m^2
June catchment solar radiation	usSolarRadWin	W/m^2
Runoff weighted June catchment solar radiation	usSolarRadWin_Q	W/m^2
Upstream Variables 2 (Topography)		
Average slope of catchment calculated from 30m DEM grid	usAveSlope	angle - degree's
Average elevation in up stream catchment	usCatElev	m
Average elevation in up stream catchment flow weighted	usCatElev_Q	m
Lake index	usLake	Dimensionless
Proportion of catchment with slope >30° (steep)	usLowGrad	%
Proportion of catchment with slope <30° (not steep)	usSteep	%
% annual runoff volume from area of catchment with slope < 30°	usLowGrad_Q	%
% annual runoff volume from area of catchment with slope > 30°	usSteep_Q	%
Upstream Variables 3 (Geology)		
% of catchment in LRI category (alluvium)	usAlluvium	%
% of catchment annual runoff from LRI category (alluvium)	usAlluvium_Q	%
% of catchment in LRI category (glacial)	usGlacial	%
% of catchment annual runoff from LRI category (glacial)	usGlacial_Q	%
% of catchment in LRI category (peat)	usPeat	%
% of catchment annual runoff from LRI category (peat)	usPeat_Q	%
Catchment average of calcium	usCalc	Ordinal scale
Runoff weighted catchment average of calcium	usCalc_Q	Ordinal scale
Catchment average of hardness (induration)	usHard	Ordinal scale
Runoff weighted catchment average of hardness (induration)	usHard_Q	Ordinal scale

Variable description	Variable Name	Units
Catchment average of particle size	usParticalSize	Ordinal scale
Runoff weighted catchment average of particle size	usParticalSize_Q	Ordinal scale
Catchment average of phosphorus	usPhos	Ordinal scale
Runoff weighted catchment average of phosphorus	usPhos_Q	Ordinal scale
Upstream Variables 4 (Land Cover)		
% of catchment in LCDB category (bare ground)	usBare	%
% of annual runoff from LCDB category (bare)	usBare_Q	%
% of catchment in LCDB category (exotic forest)	usExoticForest	%
% of annual runoff from LCDB category (exotic forest)	usExoticForest_Q	%
% of catchment in LCDB category (indigenous forest)	usIndigForest	%
% of annual runoff from LCDB category (indigenous forest)	usIndigForest_Q	%
% of catchment in LCDB category (mangrove, riparian, willows, coastal sands)	usMangrove	%
% of annual runoff from LCDB category (mangrove, riparian, willows, coastal sands)	usMangrove_Q	%
% of catchment in LCDB category (other than category 1-9)	usMiscLandCover	%
% of annual runoff from LCDB category (other than category 1–9)	usMiscLandCover_Q	%
% of annual runoff from LCDB category (pastoral)	usPastoral_Q	%
% of catchment in LCDB category (pastoral)	usPastoral	%
% of catchment in LCDB category (scrub)	usScrub	%
% of annual runoff from LCDB category (scrub)	usScrub_Q	%
% of catchment in LCDB category (tussock)	usTussock	%
% of annual runoff from LCDB category (tussock)	usTussock_Q	%
% of catchment in LCDB category (urban)	usUrban	%
% of annual runoff from LCDB category (urban)	usUrban_Q	%
% of catchment in LCDB category (inland and coastal wetlands)	usWetland	%
% of annual runoff from LCDB category (wetlands)	usWetland_Q	%