University of Southern Queensland

Faculty of Health, Engineering and Sciences

Using desktop hydrologic data to predict fish presence in streams in northern British Columbia

A Dissertation submitted by

B. Byrd

in fulfilment of the requirements of

ENG4112 Research Project

towards the degree of

Bachelor of Engineering - Honours (Environmental Engineering)

Submitted: October, 2014

Abstract

Identification of fish-bearing streams is a key part of many environmental assessments in Canada in general, and specifically in British Columbia (BC), where fish and fish habitat are highly valued components of the natural environment. Pre-field identification of likely fish-bearing and non-fish-bearing streams has the potential to reduce cost and effort related to field inventories, and to expedite the project design process.

Previous research has considered desktop level hydrologic, geologic and land-use data from single catchments with good results, but in some cases did not maintain similar predictive success for distant catchments. This research drew from three distinct catchments, with the aim of developing a model that will be more generally applicable. Data on fish presence/absence, watershed area, and mean and maximum monthly flows was collected from 2055 stream crossing points as part of the environmental assessment for the Prince Rupert Gas Transmission (PRGT) project. Canadian Digital Elevation Data was used to identify the elevation and derive the slope for each site. Parameters derived from this data were assessed using logistic regression to develop a model for predicting fish-bearing status.

The final model included the following parameters: watershed area, field gradient (as a proxy for higher-quality desktop slope values), number of months per year with maximum flow $\geq$ the 80th percentile of maximum monthly flows, and latitude. The model achieved good predictive success for non-fish-bearing streams (79% to 91% correctly identified) but performed less well for fish-bearing streams (65% to 66% correctly identified). The contrast between levels of predictive success was thought to be strongly influenced by the quality of the underlying data, where, for regulatory reasons, the actual status of streams classified as non-fish-bearing was likely far more certain than the status of streams classified as fish-bearing.

University of Southern Queensland

Faculty of Health, Engineering & Sciences

ENG4111/2 *Research Project*

Limitations of Use

Certification of Dissertation

I certify that the ideas, designs and experimental work, results, analyses and conclusions set out in this dissertation are entirely my own effort, except where otherwise indicated and acknowledged.

I further certify that the work is original and has not been previously submitted for assessment in any other course or institution, except where specifically stated.

B. Byrd

0050078399

Signature

26 October 2014

Date

Acknowledgments

I would like to acknowledge all of my colleagues at Stantec Burnaby for their enthusiasm in providing guidance (and text books) for my work. In particular, Sandra Nelson and Phil Molloy for help with statistics, Afshin Parsamanesh and Kirby Ottenbreit for help understanding the mysteries of fish surveys, and Ward Prystay for help getting access to data. However, special acknowledgement goes to my wife Anna and son Finley, who put up with 7-day work weeks on many occasions before this work was finished.

B. Byrd

University of Southern Queensland

October 2014

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Identification of fish-bearing streams, and subsequent assessment of stream habitat characteristics and potential effects on both fish and fish habitat, is a key part of many environmental assessments in Canada in general, and specifically in British Columbia (BC), where fish and fish habitat are highly valued components of the natural environment. Significant time and money is spent identifying and assessing fish-bearing streams potentially affected by projects undergoing environmental assessments.

Regulations for classifying streams as fish-bearing or non-fish-bearing are moderately strict, and require field inventories for confirmation of status. However, pre-field identification of likely fish-bearing and non-fish-bearing streams has the potential to reduce cost and effort related to field inventories, and to help expedite and streamline the project design process.

Desktop hydrologic data (i.e., available without field surveys) is often used in preliminary assessment of streams, and could potentially be used more systematically to predict for fish presence.

1.1 Research Aim

The primary aim of this research was to create a method for using desktop hydrologic data collected and analysed during environmental assessments to predict fish presence in streams in BC, for more efficient allocation of ground-truthing field work by fisheries biologists. Key objectives in reaching the aim were:

- Identification of desktop available hydrologic and related data which may correlate with fish presence
- Analysis of potential correlations to assess which parameters show correlation that is statistically significant ($\alpha < 0.05$)
- Development of a modelled parameter set for all data inputs shown to be individually significant
- Transformation of the parameter set into a predictive statistical model

Chapter 2

Background

2.1 Regulatory Context

Environmental assessments in BC under the *BC Environmental Assessment Act* (2002), and in Canada in general under the *Canadian Environmental Assessment Act, 2012* (Government of Canada 2013*a*), are based around the assessment of effects on *valued components* (VCs) (Environmental Assessment Office 2013). This approach is grounded in the work of Beanlands & Duinker (1983) about approaches to environmental impact assessments (Environmental Assessment Office 2013). Beanlands & Duinker (1983) emphasise the need to identify a set of valued ecological components (VECs) at the start of the environmental assessment process, in order to focus the assessment appropriately.

In BC, VCs are defined as "components of the natural and human environment that are considered by the proponent, public, Aboriginal groups, scientists and other technical specialists, and government agencies involved in the assessment process to have scientific, ecological, economic, social, cultural, archaeological, historical, or other importance" (Environmental Assessment Office 2013). The Canadian Environmental Assessment Agency uses Beanlands and Duiker's VEC terminology, defining VECs as "[t]he environmental element of an ecosystem that is identified as having scientific, social, cultural, economic, historical, archaeological or aesthetic importance" (Canadian Environmental Assessment Agency 2009).

Because of the importance of fish and fisheries to Canada, and particularly BC, from commercial, Aboriginal and recreational perspectives, one of the major VCs assessed in almost all environmental assessments in BC is *Fish and Fish Habitat*, sometimes subsumed under a broader VC such as *Freshwater Aquatic Resources* (Stantec Consulting Ltd. 2014). A key aspect of assessing environmental effects on this VC is a baseline assessment of the existence and location of fish habitat in streams which may be affected by a proposed project. The time and financial costs associated with the field work required to collect this baseline data can be very high; thus, any methods to make this field work more efficient and cost effective could result in substantial cost and time savings. This is especially the case for proposed projects with significant linear features, e.g., mines with road and rail alignments, or pipeline projects. These projects can have many hundreds of stream crossings, each of which need to be assessed for potential effects on fish or fish habitat. For example, for the environmental assessment for the recent Prince Rupert Gas Transmission (PRGT) project, over 800 stream crossings were part of the final pipeline alignment, and over 2000 crossing were assessed for fish-bearing status (Stantec Consulting Ltd. 2014).

The requirements for these assessments are in part because of section 35.(1) of the Canadian *Fisheries Act* (Government of Canada 2013*b*), which states that *"No person shall carry on any work, undertaking or activity that results in serious harm to fish that are part of a commercial, recreational or Aboriginal fishery, or to fish that support such a fishery."* The Fisheries Protection Policy Statement (Fisheries and Oceans Canada 2013) under the *Fisheries Act* defines "serious harm to fish" as "death of fish", or "permanent alteration" or "destruction of fish habitat". Thus, to meet the requirements of these regulations, all streams which may be affected by a project must be assessed to determine whether fish and fish habitat are present, i.e., whether the stream is fish-bearing or not.

Other key regulatory drivers for including fish and fish habitat in environmental assessments are the *Species at Risk Act* (Government of Canada 2013*c*), which under section 58.(1)(b) provides protection for listed aquatic species, and section 11(a) of the Environmental Protection and Management Regulation (2013), under the BC *Oil and Gas Activities Act* (Province of British Columbia 2008), which states that stream crossings for oil and gas activities must be constructed so that they are "unlikely to harm fish or destroy, damage or harmfully alter fish habitat".

Established under the BC *Oil and Gas Activities Act*, the BC Oil and Gas Commission provides guidance on classification of streams in their Environmental Protection and Management Guide (BC Oil and Gas Commission 2013). The guide notes that streams should be classified as types S1 through to S6, where types S1 to S4 are fish-bearing streams of varying types and widths, and types S5 to S6 are non-fish-bearing streams (BC Oil and Gas Commission 2013, Forest Service British Columbia 1998, Province of British Columbia 2013). One of the first differentiations in stream classification is determination of whether a stream is a fish stream. A fish stream is defined under the Environmental Protection and Management Regulation (BC Oil and Gas Commission 2013, Forest Service British Columbia 1998, Province of British Columbia 2013) as a stream frequented by either anadromous salmonids, rainbow trout, cutthroat trout, brown trout, bull trout, Dolly Varden char, lake trout, brook trout, kokanee, largemouth bass, smallmouth bass, mountain whitefish, lake whitefish, arctic grayling, burbot, white sturgeon, black crappie, yellow perch, walleye or northern pike, or a species identified as either at risk or regionally important. Streams are also by default classified as fish streams if they have gradients less than 20%, unless proven otherwise by an acceptable fish inventory (BC Oil and Gas Commission 2013, Forest Service British Columbia 1998).

2.2 Current Approaches

Determination of fish presence is made in accordance with methods and standards provided by the BC Resources Information Standards Committee (RISC) (formerly the BC Resource Information Committee) (BC Oil and Gas Commission 2013). The Resources Inventory Committee (RIC) Standard for Reconnaissance (1:20,000) Fish and Fish Habitat Inventories (BC Fisheries Information Services Branch 2001) sets the standard for reconnaissance level sample-based surveys covering whole watersheds. The 1:20000 reconnaissance is the basis for "intensive level inventories" required for fish stream identification (BC Oil and Gas Commission 2013, p. 1:6). The RISC standard suggests that fish stream classification (along with other objectives of fish and fish habitat inventories) begin with identification and classification of streams using maps and aerial photos. In particular, the standard suggests review of the Fisheries Information Summary System (FISS), a BC-wide data set on fish, fishing and fish habitat; recording of FISS and other desktop data in Field Data Information System (FDIS), "an MS Access data capture and reporting tool for fish and fish habitat data collected to Resource Information Standards Committee (RISC) standards" (BC Ministry of Environment n.d.); and use of the Fish and Fish Habitat Assessment Tool (FHAT20), a computer program that uses characteristics from 1:20,000 scale mapping and aerial photos to predict fish presence, along with other outputs (BC Fisheries Information Services Branch 2000, BC Fisheries Information Services Branch 2001). While FISS is commonly used for environmental assessment baseline studies, and databases based on FDIS are in use, FHAT20 is not commonly used (Parsamanesh 2014*a*, pers. comm., 2 June 2014).

Predictions of fish presence by FHAT20 seem to be based mostly on fish habitat characteristics and known fish presence in other streams (as recorded in the FDIS used as input to FHAT20). Thus, FHAT20 may not be a particularly useful tool for predicting fish presence in areas with little previous study. This is often the case for major environmental assessment projects in BC, which predominantly take place in remote northern areas of the province. This limitation accounts for the lack of use of FHAT20 within the context of environmental assessments.

FHAT20 uses a range of outputs to predict fish presence. It outputs the probability of capability for predicting fish presence. That is, it outputs the probability that a stream reach "has no capability (that the abundance is less than 1 fish in sample site area)", as well as the probabilities of low, medium and high capability. It can also provide a "Most Probable Stream Class", which would indicate fish presence for classes S1 to S4, or absence for classes S5 and S6. FHAT20 can also output "FPC Fish Presence" based on probabilities and user defined probability limits (BC Fisheries Information Services Branch 2000, pp. 16-17). While these outputs are similar to those targeted by this project, the input requirements for FHAT20 are much more detailed and site specific than the inputs used for this analysis, which targets situations where little previous field study has occurred.

Calculation of probabilities in FHAT20 are based on Gaussian multivariant kernel analysis with a "Bayesian sampling-importance-resampling algorithm" (BC Fisheries Information Services Branch 2000). The Bayesian algorithm likely uses analytical integration to eliminate "nuisance" parameters (such as the observation error variance and catchability coefficient) from probability calculations in order to reduce computational load, but that is beyond the scope of this review (BC Fisheries Information Services Branch 2000, Walters & Ludwig 1994).

As well as using FISS data when conducting initial desktop reviews of streams, fisheries biologists often work with some desktop-available hydrologic data (usually mean monthly flows and means of daily maximum flows). Hydrographs of this data are primarily used to identify suitable site visit times and assess changes in flows caused by projects, but they are also used to make preliminary judgements on productive capacity, which informs habitat classification and, potentially, fish-bearing status (Parsamanesh 2014*b*, pers. comm., 3 June 2014). However, use of this hydrologic data is not systematic, and relies more on professional experience and judgement than on a consistent, reproducible approach. While flow data may be the only data that can be derived from desktop sources for some sites, other useful hydrologic data and related desktop-available data could potentially be used for many sites.

The more systematic approach that was the aim of this project was to identify which specific aspects of desktop-available hydrologic—and other related—data provides the highest probability of correctly identifying a stream as fish-bearing, and to quantify the relative importance of specific indicators. This more systematic approach could allow focus of field programs on sites that have higher uncertainty regarding fish-bearing status. It could also assist in initial project design by early identification and elimination of routes or design options likely to affect streams with high likelihood of being fish-bearing. This could also help reduce the scope of field programs by reducing the number of alternative route or site options that would require assessment.

2.3 Previous Research

Previous research has been done to develop models for predicting fish presence (or presence-absence). Some modelling has focused on very localised predictive inputs (e.g., stream substrate, water depth, water temperature, instream cover, flow velocity) (Joy & Death 2000, Joy & Death 2002, Mastrorillo, Lek, Dauba & Belaud 1997, Mugodo, Kennard, Liston, Nichols, Linke, Norris & Lintermans 2006). This approach to modelling is not useful for the aims of this project, as it relies on detailed site-specific data, which could only be obtained by field studies; the purpose of this project was to rely on desktop-available data. Other models have considered desktop level hydrologic, geologic and land-use data from single catchments with good results (70% to over 90% correct classifications) (Filipe, Cowx & Collares-Pereira 2002, Joy & Death 2004, Porter, Rosenfeld & Parkinson 2000). However, in some cases, models did not maintain similar predictive success for distant catchments (Porter et al. 2000).

Modelling approaches often involved the use of artificial neural networks (ANN) (Joy & Death 2004, Mastrorillo et al. 1997). Logistic regression, linear discriminant analysis, classification trees and nearest-neighbour analyses have also been used (Filipe et al. 2002, Mugodo et al. 2006, Olden & Jackson 2002, Porter et al. 2000). ANN and classification tree based models tend to perform better than traditional methods (Olden & Jackson 2002).

The success of models using desktop-available data at a watershed level was promising. However, the usefulness of a modelling tool for long linear projects (such as major pipelines) that are not moderately consistent between watersheds would be limited. Also, development of a complex ANN-based model, or models of similar complexity, was considered beyond the scope of this project. However, identification of modelling inputs that are most highly influential in predicting fish presence, such as latitude and total catchment rainfall, as identified by Joy & Death (2004), could be helpful in identifying key predictive parameters.

2.4 Analysis Approaches

As noted in Section 3.3, a variety of analytic approaches have been used in related previous research. These include relatively complex models based on ANN and classification trees, and simpler numerical methods such as logistic regression. Because of its relative simplicity, and previous experience with other types of regression analysis, logistic regression analysis was used to check for potential correlations between parameters from the available data set and the fish-bearing status of streams in the data set.

Logistic regression allows regression analysis of categorical data such as the yes/no data for fish-bearing status (Quinn 2002). In fact, such dichotomous data sets (binary data) are the simplest case for using logistic regression (Gotelli 2004). Logistic regression fits an S-shaped (sigmoidal) curve to the data (in this case, fish-bearing = 1 and non-fish-bearing = 0), using a maximum likelihood (ML) approach, based on the function (where $\pi(x_i)$ is the probability of being fish-bearing) (Gotelli 2004, Quinn 2002):

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.1}$$

Fitting uses ML rather than least squares estimation, because binary data types have error terms with binomial distribution, rather than a normal distribution which is required for least squares estimation to be appropriate. For non-normal distributions, ML estimation is generally performed through iterative approaches (Quinn 2002). Modelling by logistic regression is performed by transforming the function into a linear model by a logit (also known as log-odds) transformation (Quinn 2002, Gotelli 2004, Dalgaard 2009, Whitlock 2009):

$$ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_i \tag{2.2}$$

Identifying ML seeks to maximise the likelihood function $L(\beta)$, where (Quinn 2002, Dalgaard 2009):

$$L = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{2.3}$$

For ease of calculation, maximisation of $log(L)$ is usually undertaken, rather than L (Quinn 2002).

The preferred method of fit testing of the sigmoid generated through ML estimation is using the log-likelihood ratio (sometimes referred to as deviance), $-2LL$ (also G or G_2 when defined without the negative), where (Zar 1996, Quinn 2002, Whitlock 2009, Field 2012):

$$-2LL = -2ln\left(\frac{L[\beta_0]}{L[\beta_0+\beta_1 x_1]}\right) \tag{2.4}$$

The log-likelihood ratio compares the log-likelihood of the full model, with the model case with parameters constrained to match the null hypothesis (H_0). Comparing the value of $-2LL$ with a χ^2 value with 1 degree of freedom and significance level (α) of 0.05 allows determination of whether the null hypothesis can be rejected. Where $-2LL > \chi^2_{1,\alpha=0.05}$, the null hypothesis can be rejected (Whitlock 2009). Calculations of fit parameters β_0 and β_1, and of $-2LL$ and the level of significance associated with the $-2LL$ value, are generally performed with computer statistical packages (Whitlock 2009, Field 2012).

Logistic regression and testing with the log-likelihood ratio can also be used to model the potential correlations between multiple variables (Quinn 2002). The logit transformation for a multiple logistic regression takes the form (Gotelli 2004, Quinn 2002):

$$ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_p x_{ip} \tag{2.5}$$

Testing of the multiple logistic regression is again similar to that for simple logistic regression. In this case, $-2LL$ for the overall model is calculated by (Quinn 2002, Field 2012):

$$-2LL = -2ln\left(\frac{L[\beta_0]}{L[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_p x_{ip}]}\right) \tag{2.6}$$

In addition to testing the overall model, it is also possible to test the model against a series of "reduced" models where only a single parameter (β) is eliminated from the likelihood ratio, for example, eliminating β_1 to check if this predictor makes the model better (Quinn 2002, Field 2012):

$$-2LL = -2ln\left(\frac{L[\beta_0 + \beta_2 x_{i2} ... + \beta_p x_{ip}]}{L[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_p x_{ip}]}\right) \tag{2.7}$$

Chapter 3

Data Collection and Transformation

The aim of this research was to use data available without conducting field surveys. Thus, other than fish presence data, obtained through a combination of field and desktop methods, most data used in analyses for this research were obtained without field verification. As discussed below, the single exception to this was the use of local gradient data collected during field surveys.

3.1 Data Sources

Data for this research was obtained from two data sets. Fish presence, watershed areas, mean monthly flows, maximum monthly flows and gradient data were obtained from the integrated fisheries information database developed by Stantec for the PRGT project. After elimination of sites from the database at various stages of quality control checks, site data was available for 2055 stream-crossing sites across four distinct hydrodynamic regions.

In addition to these data, 1:50000 digital elevation data (DEM) from Canadian Digital Elevation Data (CDED) was sourced from GeoBase, an initiative by various Canadian governments overseen by the Canadian Council on Geomatics.

3.2 Data Transformation

3.2.1 PRGT Data

While data for 2055 sites was available from the PRGT data set, not all sites had the same data available. Because of iterations in project design, especially in pipeline routing, the extent of hydrologic analysis and of field surveys varied greatly. Of the 2055 sites, only 653 sites had data for fish presence, watershed areas, mean monthly flows, maximum monthly flows and gradient.

However, in order to make the best use of those sites with limited data, site data was initially separated out into larger sets of all sites with each of watershed areas, mean monthly flows, maximum monthly flows and gradients. The 653 sites with all data were randomly split into two sets: one for model development (327 sites) and one for model testing (326 sites). Data availability for each type is summarised in Table 3.1.

Table 3.1: Summary of available data.

Data	**Number of Sites**
Watershed area	844
Mean monthly flows	246
Maximum monthly flows	414
Gradient	840
Digital elevation (CDED 1:50000)	2055
All data	**653**
All data (modelling set)	**327**
All data (testing set)	**326**

Hydrologic Data

In order to transform the mean and maximum monthly flow data into forms more potentially useful for further analysis, for each site with this data, the following parameters were calculated for both mean and maximum monthly flows:

- Maximum of monthly flows

- Minimum of monthly flows
- Average of monthly flows
- 5th percentile of monthly flows
- 10th percentile of monthly flows
- 20th percentile of monthly flows
- 80th percentile of monthly flows
- 90th percentile of monthly flows
- 95th percentile of monthly flows
- Number of months per year with flows $\geq$ average of monthly flows
- Number of months per year with flows $\leq$ the 5th percentile of monthly flows
- Number of months per year with flows $\leq$ the 10th percentile of monthly flows
- Number of months per year with flows $\leq$ the 20th percentile of monthly flows
- Number of months per year with flows $\geq$ the 80th percentile of monthly flows
- Number of months per year with flows $\geq$ the 90th percentile of monthly flows
- Number of months per year with flows $\geq$ the 95th percentile of monthly flows

Gradient

Gradient data available for some sites generally consisted of one to three field measurements of stream gradient at various points of the stream reach at the potential stream crossing. While this data is not desktop-available, it was included in analyses to compare with the slope data derived from the 1:50000 DEM data (see Section 3.2.2). The resolution of this DEM is reasonably coarse, but is the finest that is publicly available.

Higher resolution DEM is often available for purchase (such as 25 m pixel size DEM derived from 1:20000 BC's Terrain Resource Information Management (TRIM) data, available from GeoBC), but was not available for this analysis. Higher resolution DEM would result in slopes more indicative of local conditions at sites. The gradient data available from field surveys was averaged (where more than one measurement had been taken) and was used as a proxy for slopes derived from higher resolution DEM.

3.2.2 DEM Data

The DEM data sourced from CDED is provided as a series of raster images. Applicable map tiles with CDED were identified by overlaying the National Topographic System (NTS) grid tiles with the latitude and longitude of each of the 2055 sites in the PRGT data set in the Quantum GIS (QGIS) software package. Fifty-five DEM images were then imported and merged in QGIS.

Elevation data was extracted from the DEM for each site. Slopes at each site were then also derived from the DEM using the GDAL/DEM Slope function within QGIS. As most data was within a reasonably narrow latitudinal band (approximately 54.2 °N to 56.3 °N), z-factor conversions of latitude and longitude were used to produce elevation, rather than re-projecting the DEM. Based on an approximate latitude of 55.5 °N, z-factor was 8.8984×10^{-6}.

Chapter 4

Regression Analysis

As discussed in Section 2.3, a number of modelling approaches have previously been used to develop models whose purpose is similar to the aim of this project. This research used logistic regression to identify potential correlations between input variables and fish presence. Log-likelihood ratios, transformed into various R^2 values, were used to test the fit of models.

Multivariant logistic regression analysis was undertaken using those inputs that yielded promising correlations when assessed on an individual basis.

Logistic regression and further statistical analysis was undertaken using RStudio and the underlying R computer statistics package.

4.1 Single Logistic Regressions

To make best use of data from sites without comprehensive data, and to assist in identifying parameters with reasonable potential for predicting fish presence, individual data sets for each parameter (as discussed in Section 3.2.1) were used to carry out binomial logistic regression using the `glm` command in RStudio. In order to simplify the process of generating models for all of the individual parameters, and to produce R-statistics R^2_L (Hosmer and Lemeshow), R^2_{CS} (Cox and Snell) and R^2_N (Nagelkerke), the R function in Appendix B.1 was used.

The results of these models were used to select parameters for further modelling using multivarient logistic regression. For each parameter, values for deviance ($-2LL$), the significance of $-2LL$, and correlation measures R^2_L , R^2_{CS}, R^2_N and odds ratio were inspected. Correlation measures were calculated within the R function in Appendix B.1, as:

$$R^2_L = \frac{-2LL_{model}}{-2LL_{null}} \tag{4.1}$$

$$R^2_{CS} = 1 - exp\left(\frac{(-2LL_{model}) - (-2LL_{null})}{n}\right) \tag{4.2}$$

$$R^2_N = \frac{R^2_{CS}}{1 - exp\left(\frac{-2LL_{null}}{n}\right)} \tag{4.3}$$

Parameters were excluded from further modelling if the significance of χ^2 was greater than 0.1. This excluded the following parameters:

- Number of months per year with flows $\geq$ average of mean monthly flows
- 10th percentile of mean monthly flows
- 20th percentile of mean monthly flows
- Longitude
- 10th percentile of maximum monthly flows
- 20th percentile of maximum monthly flows

If modelling produced no results for β_1 for a given parameter, this parameter was also excluded from further analysis. This excluded the following parameters:

- Minimum of maximum monthly flows
- 5th percentile of maximum monthly flows
- Number of months per year with flows $\geq$ the 95th percentile of maximum monthly flows
- Minimum of mean monthly flows

- 5th percentile of mean monthly flows
- Number of months per year with flows $\geq$ the 95th percentile of mean monthly flows

Results of the analyses from these model runs for parameters that were carried forward are shown in Table 4.1 (ordered by descending value of R^2_N).

4.2 Multivariant Logistic Regression

In order to avoid potential errors associated with stepwise methods (Field 2012), the initial multivarient logistic regression was run by forced-entry method (i.e., all parameters were included). Results of the forced entry model run are summarised in Table 4.2. Of the 25 parameters included in the initial model, only 8 had significant z-values (i.e., were considered to contribute significantly to the model):

- Watershed area
- Average of the maximum monthly flows
- 80th percentile of maximum monthly flows
- Months with flows $\leq$ the 20th percentile of maximum monthly flows
- Average of the mean monthly flows
- Months with flows $\leq$ the 5th percentile of maximum monthly flows
- Months with flows $\leq$ the 10th percentile of maximum monthly flows

However, 2 additional parameters were also close to the significance threshold ($\alpha = 0.05$):

- Gradient
- Latitude

The R function in Appendix B.2 was used to produce key statistics about the model. In addition to the previously noted correlation measures and odds ratio, this included

Table 4.1: Results of single logistic regressions.

Input Parameter	β_0	β_1	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**Odds Ratio**
Watershed area	-1.55480	0.08732	94.40212	0.00000	0.10610	0.10594	0.16250	1.09125
Gradient	0.49502	-0.08786	73.14326	0.00000	0.06296	0.08339	0.11131	0.91589
Maximum (max. flows)	-1.97514	5.45530	23.17979	0.00000	0.06167	0.05458	0.09134	233.995
95th %ile (max. flows)	-1.97163	7.07284	21.70626	0.00000	0.05775	0.05120	0.08569	1179.49
Average (max. flows)	-2.09248	22.34291	19.89573	0.00001	0.05293	0.04703	0.07871	5.0513×10^9
90th %ile (max. flows)	-1.95827	9.36159	19.16185	0.00001	0.05098	0.04534	0.07587	11632.9
80th %ile (max. flows)	-1.95568	12.82130	18.06237	0.00002	0.04805	0.04279	0.07161	370016
95th %ile (mean flows)	-1.64458	21.03676	10.71251	0.00106	0.03702	0.04278	0.06173	1.3682×10^9
Maximum (mean flows)	-1.62570	18.48484	10.52107	0.00118	0.03635	0.04203	0.06065	1.0663×10^8
90th %ile (mean flows)	-1.63793	23.68672	10.46430	0.00122	0.03616	0.04181	0.06033	1.9365×10^{10}
Months with flows $\geq$ 80th %ile (max. flows)	-0.86958	-0.12161	14.18136	0.00017	0.03773	0.03375	0.05649	0.88549
Months with flows $\leq$ 20th %ile (max. flows)	-0.41414	-0.15807	12.05298	0.00052	0.03206	0.02876	0.04813	0.85379
Average (mean flows)	-1.82484	53.17985	8.10457	0.00442	0.02800	0.03254	0.04695	1.2466×10^{23}
Months with flows $\leq$ 5th %ile (max. flows)	-0.48992	-0.14863	11.39638	0.00074	0.03032	0.02722	0.04555	0.86189
Months with flows $\leq$ 10th %ile (max. flows)	-0.48992	-0.14863	11.39638	0.00074	0.03032	0.02722	0.04555	0.86189
Months with flows $\geq$ 90th %ile (max. flows)	-1.13437	-0.09941	10.15638	0.00144	0.02702	0.02429	0.04065	0.90537
Months with flows $\geq$ average (max. flows)	-2.34878	0.26702	10.04765	0.00153	0.02673	0.02403	0.04022	1.30607
80th %ile (mean flows)	-1.44025	25.79398	6.66566	0.00983	0.02303	0.02684	0.03872	1.5929×10^{11}
Latitude	43.08583	-0.79425	34.60936	0.00000	0.01392	0.01668	0.02378	0.45192
Months with flows $\leq$ 5th %ile (mean flows)	-0.32192	-0.08863	3.66989	0.05540	0.01268	0.01487	0.02145	0.91519
Months with flows $\leq$ 10th %ile (mean flows)	-0.32192	-0.08863	3.66989	0.05540	0.01268	0.01487	0.02145	0.91519
Months with flows $\geq$ 80th %ile (mean flows)	-0.58230	-0.05801	3.11316	0.07766	0.01076	0.01263	0.01822	0.94364
Months with flows $\leq$ 20th %ile (mean flows)	-0.33541	-0.08535	3.04979	0.08075	0.01054	0.01237	0.01785	0.91819
Elevation	-0.62680	-0.00057	11.67427	0.00063	0.00596	0.00673	0.00993	0.99943
Slope	-0.69883	-0.01627	11.50528	0.00069	0.00463	0.00558	0.00796	0.98387

Table 4.2: Results of multivarient logistic regression model 1.

	β_0	**Std. Error**	**z-value**	**Signif.**
Intercept	54.8714	27.9673	1.962	0.0498
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.696	0.2567	2.7115	0.0067
Gradient	-0.0425	0.0246	-1.7292	0.0838
Maximum (max. flows)	-1156.0209	832.4499	-1.3887	0.1649
95th %ile of (max. flows)	2671.5608	1863.4778	1.4336	0.1517
Average (max. flows)	-639.9905	271.0544	-2.3611	0.0182
90th %ile (max. flows)	-1664.0722	1120.0362	-1.4857	0.1374
80th %ile (max. flows)	476.1183	216.2012	2.2022	0.0277
95th %ile (mean flows)	4799.5059	6143.6564	0.7812	0.4347
Maximum (mean flows)	-2410.2925	2809.7857	-0.8578	0.391
90th %ile (mean flows)	-2822.6687	3705.3696	-0.7618	0.4462
Months with flows $\geq$ 80th %ile (max. flows)	-0.0821	0.0985	-0.8339	0.4043
Months with flows $\leq$ 20th %ile (max. flows)	-3.5911	1.434	-2.5043	0.0123
Average (mean flows)	1261.9928	527.7316	2.3914	0.0168
Months with flows $\leq$ 5th %ile (max. flows)	-3.2243	1.2145	-2.6549	0.0079
Months with flows $\leq$ 10th %ile (max. flows)	6.8749	2.395	2.8705	0.0041
Months with flows $\geq$ 90th %ile (max. flows)	20.0248	0.0776	0.32	0.749
Months with flows $\geq$ average (max. flows)	-0.0767	0.0539	-1.4228	0.1548
80th %ile (mean flows)	-30.0139	482.7067	-0.0622	0.9504
Latitude	-0.9986	0.5202	-1.9198	0.0549
Months with flows $\leq$ 5th %ile (mean flows)	-0.1935	0.974	-0.1987	0.8425
Months with flows $\leq$ 10th %ile (mean flows)	0.1684	1.7952	0.0938	0.9252
Months with flows $\geq$ 80th %ile (mean flows)	0.0428	0.0993	0.4306	0.6667
Months with flows $\leq$ 20th %ile (mean flows)	-0.5233	1.1496	-0.4552	0.649
Elevation	0.0011	0.0007	1.4928	0.1355
Slope	-0.0057	0.0188	-0.3024	0.7623

the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). Key statistics for this model are summarised in Table 4.3.

Table 4.3: Summary statistics for model 1 (comparison with H_0).

Model no.	$-2LL$	**Signif.**	R_L^2	R_{CS}^2	R_N^2	**AIC**	**BIC**	**Odds Ratio**
1	146.7	0.000	0.3236	0.3615	0.4820	356.6	596.1	2.006

As an initial check of the first model, fitted values from the model were used to predict fish presence for the both modelling data set, and as a validation check, against the testing data set. Results of the check are summarized in Table 4.4, where *sensitivity* refers to the proportion of true positives (i.e., correctly identified fish-bearing steams), *specificity* refers to the proportion of true negatives (i.e., correctly identified non-fish-bearing streams), *PPV (positive prediction value)* refers to the proportion of positive that are true (i.e., proportion of streams correctly identified as fish-bearing out all all streams identified as fish-bearing), *NPV (negative prediction value)* refers to the proportion of negatives that are true (i.e., proportion of streams correctly identified as non-fish-bearing out of all streams identified as non-fish-bearing), *accuracy* refers to the overall proportion of correct predictions (i.e., streams correctly identified as either fish-bearing or non-fish bearing), and where *MCC (Matthews correlation coefficient)* is a generalised measure of predictive success for binary systems.

Table 4.4: Predictive performance of model 1.

Model no.	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**	**MCC**
Modelling data						
1	64.81	87.27	83.33	71.64	76.15	0.5351
Testing data						
1	58.28	86.50	81.20	67.46	72.39	0.4668

4.2.1 Parameter Refinement

Checking for Multicollinearity

Although the initial model indicated some potentially useful parameters, there was a strong suspicion that multicollinearity could be substantially affecting the model. There should have been a reasonably strong correlation between the gradient and slope parameters. And, from a hydrologic perspective, there should be some correlation between watershed area and a number of the parameters related to high flows, such as many of the parameters derived from maximum monthly flow data.

In order to identify potential correlations between parameters, an analysis of Pearson's correlation coefficient (r) was conducted for each pair of parameters in the model. Results of the analysis are presented in Table 4.5, sorted in order of descending values of R^2_N based on the analysis of the initial model. Working from the top of the table (highest values of R^2_N), parameters were eliminated from further modelling if they had $r < 0.8$ with a parameter higher on the table. This process eliminated 17 parameters. Additionally, although the slope parameter (derived from the CDED 1:50000 elevation data) showed only a moderate correlation with the gradient parameter (derived from field gradient measurements and used as a proxy for slope values from higher resolution elevation data)($r = 0.575$), it was also eliminated from the model on the basis of being a less accurate and less useful version of the same information.

Field (2012) suggests that this approach of identifying multicollinearity can miss its more subtle forms, and suggests diagnosis with variance inflation factors (VIF). Initial inspections of VIFs did indicate concern with a number of parameters (i.e., VIFs well above 10—suggested as a threshold for concern by Field (2012), citing Myers (1990)). However, the large number of parameters meant that VIFs were not useful in identifying which parameters were strongly correlated. To check if any multicollinearity existed in the remaining parameters used for model 2, VIFs were recalculated. In this case, the highest VIF was approximately 3—well below the threshold for concern. However, Field (2012) also references Bowerman & O'Connell (1990) in suggesting that if the average VIF exceeds 1, the model may be biased by multicollinearity. As the average VIF was approximately 1.73, this bias may exist.

After elimination of highly correlated parameters, the model was refined based on these

Table 4.5: Pearson's correlation coefficient (r) for all parameter pairs in the initial model.

	Watershed area	Gradient	Maximum (max. flows)	95th %ile (max. flows)	Average (max. flows)	90th %ile (max. flows)	80th %ile (max. flows)	95th %ile (mean flows)	Maximum (mean flows)	90th %ile (mean flows)	Months with flows $\geq$ 80th %ile (max. flows)	Months with flows $\leq$ 20th %ile (max. flows)	Average (mean flows)	Months with flows $\leq$ 5th %ile (max. flows)	Months with flows $\leq$ 10th %ile (max. flows)	Months with flows $\geq$ 90th %ile (max. flows)	Months with flows $\geq$ average (max. flows)	80th %ile (mean flows)	Latitude	Months with flows $\leq$ 5th %ile (mean flows)	Months with flows $\leq$ 10th %ile (mean flows)	Months with flows $\geq$ 80th %ile (mean flows)	Months with flows $\leq$ 20th %ile (mean flows)	Elevation	Slope
Watershed area	1.000	-0.080	0.981	0.991	0.943	0.995	0.894	0.984	0.970	0.990	-0.062	-0.092	0.947	-0.103	-0.098	-0.057	-0.021	0.876	0.025	-0.127	-0.121	-0.097	-0.117	-0.059	-0.093
Gradient	-0.080	1.000	-0.082	-0.080	-0.086	-0.079	-0.090	-0.079	-0.081	-0.078	0.142	0.190	-0.083	0.191	0.192	0.153	0.007	-0.085	-0.156	0.181	0.177	0.149	0.177	-0.125	0.575
Maximum (max. flows)	0.981	-0.082	1.000	0.998	0.989	0.994	0.963	1.000	0.998	0.998	-0.068	-0.101	0.991	-0.113	-0.107	-0.062	-0.013	0.952	0.030	-0.140	-0.133	-0.107	-0.128	-0.095	-0.093
95th %ile (max. flows)	0.991	-0.080	0.998	1.000	0.978	0.999	0.944	0.999	0.993	1.000	-0.066	-0.098	0.980	-0.110	-0.104	-0.060	-0.015	0.930	0.029	-0.136	-0.129	-0.104	-0.124	-0.086	-0.092
Average (max. flows)	0.943	-0.086	0.989	0.978	1.000	0.968	0.992	0.987	0.995	0.979	-0.074	-0.111	1.000	-0.124	-0.117	-0.067	-0.007	0.986	0.032	-0.153	-0.146	-0.117	-0.140	-0.119	-0.095
90th %ile (max. flows)	0.995	-0.079	0.994	0.999	0.968	1.000	0.929	0.996	0.987	0.999	-0.065	-0.096	0.971	-0.107	-0.102	-0.058	-0.016	0.914	0.029	-0.133	-0.126	-0.101	-0.122	-0.081	-0.091
80th %ile (max. flows)	0.894	-0.090	0.963	0.944	0.992	0.929	1.000	0.958	0.975	0.945	-0.079	-0.117	0.990	-0.131	-0.124	-0.071	-0.002	0.998	0.037	-0.162	-0.154	-0.124	-0.148	-0.132	-0.097
95th %ile (mean flows)	0.984	-0.079	1.000	0.999	0.987	0.996	0.958	1.000	0.998	0.999	-0.064	-0.095	0.989	-0.106	-0.101	-0.058	-0.014	0.947	0.028	-0.131	-0.125	-0.100	-0.120	-0.090	-0.090
Maximum (mean flows)	0.970	-0.081	0.998	0.993	0.995	0.987	0.975	0.998	1.000	0.994	-0.066	-0.098	0.997	-0.110	-0.104	-0.060	-0.011	0.967	0.029	-0.136	-0.130	-0.104	-0.125	-0.098	-0.091
90th %ile (mean flows)	0.990	-0.078	0.998	1.000	0.979	0.999	0.945	0.999	0.994	1.000	-0.062	-0.092	0.982	-0.103	-0.098	-0.056	-0.015	0.933	0.028	-0.128	-0.122	-0.097	-0.117	-0.084	-0.089
Months with flows $\geq$ 80th %ile (max. flows)	-0.062	0.142	-0.068	-0.066	-0.074	-0.065	-0.079	-0.064	-0.066	-0.062	1.000	0.883	-0.069	0.838	0.860	0.905	0.266	-0.072	0.117	0.648	0.660	0.636	0.674	0.402	0.161
Months with flows $\leq$ 20th %ile (max. flows)	-0.092	0.190	-0.101	-0.098	-0.111	-0.096	-0.117	-0.095	-0.098	-0.092	0.883	1.000	-0.103	0.990	0.997	0.832	0.241	-0.107	0.173	0.870	0.880	0.849	0.891	0.477	0.199
Average (mean flows)	0.947	-0.083	0.991	0.980	1.000	0.971	0.990	0.989	0.997	0.982	-0.069	-0.103	1.000	-0.116	-0.109	-0.063	-0.008	0.984	0.030	-0.143	-0.136	-0.109	-0.131	-0.111	-0.091
Months with flows $\leq$ 5th %ile (max. flows)	-0.103	0.191	-0.113	-0.110	-0.124	-0.107	-0.131	-0.106	-0.110	-0.103	0.838	0.990	-0.116	1.000	0.997	0.786	0.199	-0.120	0.185	0.922	0.929	0.875	0.935	0.501	0.196
Months with flows $\leq$ 10th %ile (max. flows)	-0.098	0.192	-0.107	-0.104	-0.117	-0.102	-0.124	-0.101	-0.104	-0.098	0.860	0.997	-0.109	0.997	1.000	0.808	0.218	-0.113	0.180	0.900	0.909	0.866	0.918	0.489	0.198
Months with flows $\geq$ the 90th %ile (max. flows)	-0.057	0.153	-0.062	-0.060	-0.067	-0.058	-0.071	-0.058	-0.060	-0.056	0.905	0.832	-0.063	0.786	0.808	1.000	0.364	-0.065	0.116	0.590	0.601	0.575	0.614	0.368	0.164
Months with flows $\geq$ average (max. flows)	-0.021	0.007	-0.013	-0.015	-0.007	-0.016	-0.002	-0.014	-0.011	-0.015	0.266	0.241	-0.008	0.199	0.218	0.364	1.000	-0.001	0.055	0.043	0.048	0.029	0.053	0.011	0.068
80th %ile (mean flows)	0.876	-0.085	0.952	0.930	0.986	0.914	0.998	0.947	0.967	0.933	-0.072	-0.107	0.984	-0.120	-0.113	-0.065	-0.001	1.000	0.034	-0.148	-0.141	-0.113	-0.136	-0.129	-0.090
Latitude	0.025	-0.156	0.030	0.029	0.032	0.029	0.037	0.028	0.029	0.028	0.117	0.173	0.030	0.185	0.180	0.116	0.055	0.034	1.000	0.188	0.194	0.148	0.193	0.417	-0.071
Months with flows $\leq$ 5th %ile (mean flows)	-0.127	0.181	-0.140	-0.136	-0.153	-0.133	-0.162	-0.131	-0.136	-0.128	0.648	0.870	-0.143	0.922	0.900	0.590	0.043	-0.148	0.188	1.000	0.998	0.915	0.994	0.497	0.167
Months with flows $\leq$ 10th %ile (mean flows)	-0.121	0.177	-0.133	-0.129	-0.146	-0.126	-0.154	-0.125	-0.130	-0.122	0.660	0.880	-0.136	0.929	0.909	0.601	0.048	-0.141	0.194	0.998	1.000	0.925	0.998	0.497	0.170
Months with flows $\geq$ 80th %ile (mean flows)	-0.097	0.149	-0.107	-0.104	-0.117	-0.101	-0.124	-0.100	-0.104	-0.097	0.636	0.849	-0.109	0.875	0.866	0.575	0.029	-0.113	0.148	0.915	0.925	1.000	0.936	0.462	0.157
Months with flows $\leq$ 20th %ile (mean flows)	-0.117	0.177	-0.128	-0.124	-0.140	-0.122	-0.148	-0.120	-0.125	-0.117	0.674	0.891	-0.131	0.935	0.918	0.614	0.053	-0.136	0.193	0.994	0.998	0.936	1.000	0.494	0.176
Elevation	-0.059	-0.125	-0.095	-0.086	-0.119	-0.081	-0.132	-0.090	-0.098	-0.084	0.402	0.477	-0.111	0.501	0.489	0.368	0.011	-0.129	0.417	0.497	0.497	0.462	0.494	1.000	-0.088
Slope	-0.093	0.575	-0.093	-0.092	-0.095	-0.091	-0.097	-0.090	-0.091	-0.089	0.161	0.199	-0.091	0.196	0.198	0.164	0.068	-0.090	-0.071	0.167	0.170	0.157	0.176	-0.088	1.000

remaining parameters:

- Watershed area
- Gradient
- Months with flows $\geq$ the 80th percentile of maximum monthly flows
- Number of months per year with flows $\geq$ average of maximum monthly flows
- Latitude
- Months with flows $\leq$ the 5th percentile of mean monthly flows
- Elevation

This second model version had worse fit, according to all R^2 values (e.g., R^2_N fell from 0.482 to 0.399), but was better by information theoretic (IT) criteria (i.e., both AIC and BIC fell). Summaries of these criteria are provided in Table 4.7. Results of the comparison of model 2 with model 1 (see Table 4.8) confirm that the reduction in goodness in fit is because those parameters eliminated to avoid multicollinearity did contribute significantly to the model.

Values for β in model 2 (see Table 4.6) were broadly similar to those in model 1 (see Table 4.2), indicating fairly stable measures of effect. Significance for most individual parameters was generally better than in model 1. However, while the *Watershed area* parameter was slightly less significant (from 0.0067 to 0.0207), significance for the already not-significant parameter *Months with flows $\geq$ the average of maximum monthly flows* worsened (0.155 to 0.208), and, similarly, significance of *Elevation* worsened from 0.135 to 0.232. Significance of the estimated intercept also worsened (0.050 to 0.083). Predictive performance (see Table 4.9) was mixed in comparison to model 1: overall accuracy decreased for the modelling data set but increased for the testing data set.

Refinement by Backwards Stepwise Method

While model refinement through stepwise approaches is generally discouraged (Field 2012, Whittingham, Stephens, Bradbury & Freckleton 2006), the highly not-significant nature of two parameters in model 2 suggested that further model refinement by the

Table 4.6: Results of multivarient logistic regression model 2.

	β_0	Std. Error	z-value	Signif.
Intercept	36.0724	20.7861	1.735	0.0827
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.1903	0.0823	2.314	0.0207
Gradient	-0.0394	0.0188	-2.096	0.0361
Months with flows $\geq$ 80th %ile (max. flows)	-0.0756	0.0424	-1.784	0.0744
Months with flows $\geq$ average (max. flows)	-0.0600	0.0477	-1.259	0.2081
Latitude	-0.6405	0.3790	-1.690	0.0910
Months with flows $\leq$ 5th %ile (mean flows)	-0.0925	0.0539	-1.715	0.0863
Elevation	0.0006598	0.0005518	1.196	0.2318

Table 4.7: Summary statistics for model 2 (comparison with H_0).

Model no.	$-2LL$	Signif.	R^2_L	R^2_{CS}	R^2_N	AIC	BIC	Odds Ratio
2	116.4	0.000	0.2568	0.2995	0.3994	350.9	417.9	1.210

Table 4.8: Summary statistics for reduced model 2 (comparison with model 1).

Model no.	$-2LL$	Signif.	R^2_L	R^2_{CS}	R^2_N	AIC	BIC	Odds Ratio
2	30.27	0.0349	0.0899	0.0884	0.1375	342.6	515.0	2.0057

Table 4.9: Predictive performance of model 2 compared with model 1.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
1	64.81	87.27	83.33	71.64	76.15	0.5351
2	70.37	77.58	75.50	72.73	74.01	0.4808
Testing data						
1	58.28	86.50	81.20	67.46	72.39	0.4668
2	70.55	87.12	84.56	74.74	78.83	0.5848

backwards stepwise method was warranted. Field (2012) suggests that the backwards method is less problematic than the forward method, especially when seeking only to fit a model, and not establish causality. As this was the case, the backwards stepwise method was used for further refinement.

Whittingham et al. (2006) indicate that some of the concern with using stepwise methods is the reliance solely on the significance of predictive parameters. In order to at least partially address these concerns, the effects of parameter removal from the model were assessed by examining parameter significance, changes in goodness of fit indicators (R^2_L, R^2_{CS} and R^2_N), and changes in IT criteria (AIC and BIC). These indicators and criteria were examined within the context of model comparison with the H_0, and comparing reduced models with original models.

Third round model refinement: The next step in refining the model was to check which of *Months with flows ≥ the average of maximum monthly flows* and *Elevation* were best removed from the model. Model 3a was created by removing the least significant *Elevation* parameter. Model 3a resulted in very slight decreases in all three R^2 indicators, but slight increases in both AIC and BIC (see Table 4.12). Changes in significance for the remaining parameters varied, with some better and others worse (see Table 4.10).

Table 4.10: Results of multivarient logistic regression model 3a.

	$\boldsymbol{\beta_0}$	**Std. Error**	**z-value**	**Signif.**
Intercept	28.2702	19.5665	1.445	0.1485
Input Parameter	$\boldsymbol{\beta_1}$	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2270	0.0840	2.701	0.0069
Gradient	-0.0425	0.0186	-2.287	0.0222
Months with flows ≥ 80th %ile (max. flows)	-0.0680	0.0417	-1.632	0.1026
Months with flows ≥ average (max. flows)	-0.0640	0.0478	-1.338	0.1810
Latitude	-0.4957	0.3561	-1.392	0.1639
Months with flows ≤ 5th %ile (mean flows)	-0.0590	0.0466	-1.267	0.2053

Model 3b was created by removing the *Months with flows ≥ the average of maximum monthly flows* parameter from model 2. This resulted in very slightly higher decreases in all three R^2 indicators than model 3a, and lower increases in both AIC and BIC (see Table 4.12). Significance for the remaining parameters was similar or better for most parameters (see Table 4.11). Significance was substantially better in model 3b compared to model 3a for the intercept (0.065 versus 0.149), as well as being better than in model 2 (0.065 versus 0.083).

Table 4.11: Results of multivarient logistic regression model 3b.

	β_0	**Std. Error**	**z-value**	**Signif.**
Intercept	38.1836	20.6840	1.846	0.0649
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.1935	0.0834	2.321	0.0203
Gradient	-0.0382	0.0186	-2.049	0.0404
Months with flows $\geq$ 80th %ile (max. flows)	-0.0924	0.0409	-2.263	0.0237
Latitude	-0.6842	0.3768	-1.816	0.0694
Months with flows $\leq$ 5th %ile (mean flows)	-0.0821	0.0534	-1.539	0.1239
Elevation	0.0007064	0.0005498	1.285	0.1988

A summary of statistics from comparing reduced models 3a and 3b with model 2 is provided in Table 4.13. The lesser significance and higher R^2 values for the model 3b comparisons indicated that the contribution provided by the *Months with flows* $\geq$ *the average of maximum monthly flows* parameter is marginally more useful in the model than that of the *Elevation* parameter.

Table 4.12: Summary statistics for models 3a and 3b (comparison with H_0).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
3a	115.0	0.000	0.2537	0.2965	0.3953	350.3	407.8	1.255
3b	114.8	0.000	0.2532	0.2960	0.3947	350.5	408.0	1.214

Table 4.13: Summary statistics for reduced models 3a and 3b (comparison with model 2).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
3a	1.429	0.2319	0.004224	0.004361	0.006765	338.8	348.5	1.210
3b	1.644	0.1997	0.004858	0.005016	0.007779	338.9	348.5	1.210

As shown in Table 4.14, predictive performance was better than model 2 for non-fish-bearing streams (specificity) for both model 3a and model 3b. Both models were worse for fish-bearing streams (sensitivity), and slightly worse overall (accuracy).

Table 4.14: Predictive performance of round 3 models compared with model 2.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
2	70.37	77.58	75.50	72.73	74.01	0.4808
3a	67.90	78.18	75.34	71.27	73.09	0.4635
3b	69.75	78.18	75.84	72.47	74.01	0.4812
Testing data						
2	70.55	87.12	84.56	74.74	78.83	0.5848
3a	69.33	87.73	84.96	74.09	78.53	0.5805
3b	67.48	87.12	83.97	72.82	77.30	0.5568

Fourth round model refinement: The next round of model refinements considered the removal of additional parameters from the model. As differences between model 3a and model 3b were marginal at best, both models were carried forward as the basis of the next set of models. Fourth round models were generated by eliminating, singly, each of the parameters in models 3a and 3b whose contribution to the models was not significant ($\alpha > 0.05$). The models were created as follows:

- Model 4a1: excluding the *Elevation* and *Months with flows* $\geq$ *the average of maximum monthly flows* parameters (see Table 4.15).
- Model 4a2: excluding the *Elevation* and *Months with flows* $\leq$ *5th percentile of mean monthly flows* parameters (see Table 4.16).
- Model 4a3: excluding the *Elevation* and *Latitude* parameters (see Table 4.17).
- Model 4a4: excluding the *Elevation* and *Months with flows* $\geq$ *80th percentile of maximum monthly flows* parameters (see Table 4.18).
- Model 4b2: excluding the *Months with flows* $\geq$ *the average of maximum monthly flows* and *Months with flows* $\leq$ *5th percentile of mean monthly flows* parameters (see Table 4.19).
- Model 4b3: excluding the *Months with flows* $\geq$ *the average of maximum monthly flows* and *Latitude* parameters (see Table 4.20).

Table 4.15: Results of multivarient logistic regression model 4a1.

	β_0	Std. Error	z-value	Signif.
Intercept	29.9496	19.4873	1.537	0.1243
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2346	0.0852	2.753	0.0059
Gradient	-0.0413	0.0184	-2.242	0.0250
Months with flows $\geq$ 80th %ile (max. flows)	-0.0850	0.0403	-2.110	0.0349
Latitude	-0.5319	0.3544	-1.501	0.1334
Months with flows $\leq$ 5th %ile (mean flows)	-0.0452	0.0455	-0.993	0.3208

Table 4.16: Results of multivarient logistic regression model 4a2.

	β_0	Std. Error	z-value	Signif.
Intercept	35.7065	18.5434	1.926	0.0542
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2853	0.0762	3.745	0.0002
Gradient	-0.0413	0.0185	-2.236	0.0253
Months with flows $\geq$ 80th %ile (max. flows)	-0.0935	0.0366	-2.555	0.0106
Months with flows $\geq$ average (max. flows)	-0.0507	0.0460	-1.102	0.2705
Latitude	-0.6390	0.3355	-1.905	0.0568

Table 4.17: Results of multivarient logistic regression model 4a3.

	β_0	Std. Error	z-value	Signif.
Intercept	1.0499	0.5083	2.066	0.0389
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2031	0.0792	2.565	0.0103
Gradient	-0.0375	0.0181	-2.070	0.0384
Months with flows $\geq$ 80th %ile (max. flows)	-0.0649	0.0413	-1.571	0.1162
Months with flows $\geq$ average (max. flows)	-0.0691	0.0480	-1.440	0.1499
Months with flows $\leq$ 5th %ile (mean flows)	-0.0800	0.0439	-1.821	0.0686

Table 4.18: Results of multivarient logistic regression model 4a4.

	β_0	Std. Error	z-value	Signif.
Intercept	26.7935	19.5593	1.370	0.1707
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2313	0.0867	2.669	0.0076
Gradient	-0.0433	0.0185	-2.336	0.0195
Months with flows $\geq$ average (max. flows)	-0.0816	0.0444	-1.839	0.0659
Months with flows $\leq$ 5th %ile (mean flows)	-0.0975	0.0407	-2.395	0.0166
Latitude	-0.4689	0.3560	-1.317	0.1878

Table 4.19: Results of multivarient logistic regression model 4b2.

	β_0	Std. Error	z-value	Signif.
Intercept	40.2270	20.3945	1.972	0.0486
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2773	0.0758	3.656	0.0003
Gradient	-0.0393	0.0185	-2.120	0.0340
Months with flows $\geq$ 80th %ile (max. flows)	-0.1113	0.0389	-2.860	0.0042
Latitude	-0.7266	0.3712	-1.958	0.0503
Elevation	0.0003	0.0005	0.543	0.5874

Table 4.20: Results of multivarient logistic regression model 4b3.

	β_0	Std. Error	z-value	Signif.
Intercept	0.6407	0.4558	1.406	0.1598
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.1823	0.0827	2.205	0.0274
Gradient	-0.0336	0.0182	-1.844	0.0652
Months with flows $\geq$ 80th %ile (max. flows)	-0.0860	0.0404	-2.129	0.0332
Months with flows $\leq$ 5th %ile (mean flows)	-0.0911	0.0532	-1.712	0.0868
Elevation	0.0004	0.0005	0.763	0.4457

For ease of comparison, significance values for parameters in the fourth round models are summarised in Table 4.21. Summary statistics for all round four models compared with the null hypothesis are summarised in Table 4.22. Comparisons between model 3a and the reduced 4a models are provided in Table 4.23. Comparisons between model 3b and the reduced 4b models are provided in Table 4.24.

Table 4.21: Significance for parameters in fourth round models.

	4a1	**4a2**	**4a3**	**4a4**	**4b2**	**4b3**
Intercept	0.1243	0.0542	0.0389	0.1707	0.0486	0.1598
Input Parameter	**4a1**	**4a2**	**4a3**	**4a4**	**4b2**	**4b3**
Watershed area	0.0059	0.0002	0.0103	0.0076	0.0003	0.0274
Gradient	0.0250	0.0253	0.0384	0.0195	0.0340	0.0652
Months with flows $\geq$ 80th %ile (max. flows)	0.0349	0.0106	0.1162	-	0.0042	0.0332
Months with flows $\geq$ average (max. flows)	-	0.2705	0.1499	0.0659	-	-
Latitude	0.1334	0.0568	-	0.1878	0.0503	-
Months with flows $\leq$ 5th %ile (mean flows)	0.3208	-	0.0686	0.0166	-	0.0868
Elevation	-	-	-	-	0.5874	0.4457

Table 4.22: Summary statistics for fourth round model versions (comparison with H_0).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
4a1	113.1	0.000	0.2495	0.2924	0.3899	350.2	398.1	1.264
4a2	113.4	0.000	0.2501	0.2930	0.3907	349.9	397.8	1.330
4a3	113.0	0.000	0.2494	0.2923	0.3897	350.2	398.1	1.225
4a4	112.3	0.000	0.2477	0.2906	0.3875	351.0	398.9	1.260
4b2	112.4	0.000	0.2480	0.2909	0.3879	350.9	398.8	1.320
4b3	111.4	0.000	0.2459	0.2888	0.3851	351.8	399.7	1.200

Of the five fourth round models, three seemed to perform better from the perspective of parameter significance. Models 4a2, 4a4 and 4b2 each had just one parameter well above significance, with another very close to $\alpha \leq 0.05$. Eliminating the least significant parameter for both 4a2 and 4b2 resulted in the same modelling set. Testing of each round four model against the null hypothesis yielded very consistent results. Model 4a2 had the highest values for all R^2 coefficients, and the lowest AIC and BIC. Of the other two models with best performing parameter significance, 4b2 was in the middle of the set, while 4a4 was consistently the second worst. However, differences in the R^2 coefficients, AIC and BIC were quite small across all models.

Table 4.23: Summary statistics for reduced models 4a1 to 4a4 (comparison with model 3a).

Model no.	$2LL$	Signif.	R_L^2	R_{CS}^2	R_N^2	AIC	BIC	Odds Ratio
4a1	1.865	0.1720	0.0055	0.0057	0.0088	340.3	349.9	1.255
4a2	1.604	0.2053	0.0047	0.0049	0.0076	340.3	349.9	1.255
4a3	1.939	0.1638	0.0057	0.0059	0.0091	340.3	349.9	1.255
4a4	2.708	0.0998	0.0079	0.0082	0.0127	340.3	349.9	1.255

Table 4.24: Summary statistics for reduced models 4b2 and 4b3 (comparison with model 3b).

Model no.	$-2LL$	Signif.	R_L^2	R_{CS}^2	R_N^2	AIC	BIC	Odds Ratio
4b2	2.342	0.1259	0.0069	0.0071	0.0110	340.5	350.1	1.214
4b3	3.324	0.0683	0.0097	0.0101	0.0156	340.5	350.1	1.214

Comparison of the reduced models against their predecessors (models 3a and 3b) (see Tables 4.23 and 4.24) supported the implications of null hypothesis testing. The parameter eliminated in model 4a2 showed least significance (0.2053), and lowest R^2 values (e.g., $R_N^2 = 0.0076$).

Predictive performance was checked for each model, working in both the modelling data set, and the testing data set. Results are summarised in Table 4.25. Within the modelling data set, model 4a1 performed the best, though models 4a2, 4b2 and 4b3 were only slightly worse. Within the testing data set, model 4b2 performed the best, closely followed by 4a2.

Based on the analysis above, model 4a2 (and 4b2) were carried forward.

Table 4.25: Predictive performance of fourth round models.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
4a1	67.28	80.00	76.76	71.35	73.70	0.4770
4a2	66.05	78.18	74.83	70.11	72.17	0.4458
4a3	64.20	76.97	73.24	68.65	70.64	0.4153
4a4	62.35	79.39	74.81	68.23	70.95	0.4239
4b2	65.43	79.39	75.71	70.05	72.48	0.4529
4b3	66.05	78.79	75.35	70.27	72.48	0.4523
Testing data						
4a1	66.26	87.12	83.72	72.08	76.69	0.5457
4a2	66.87	92.02	89.34	73.53	79.45	0.6085
4a3	63.80	87.73	83.87	70.79	75.77	0.5308
4a4	64.42	88.34	84.68	71.29	76.38	0.5434
4b2	67.48	92.64	90.16	74.02	80.06	0.6212
4b3	65.03	85.89	82.17	71.07	75.46	0.5207

Fifth round model refinement: Model 5 was a reduced version of both model 4a2 and model 4b2, including the following parameters:

- Watershed area
- Gradient
- Months with flows ≥ the 80th percentile of maximum monthly flows
- Latitude

The results of model 5 are provided in Table 4.26. Results of the comparison with the null hypothesis are summarised in Table 4.27, and comparisons with predecessor models 4a2 and 4b2 are provided in Table 4.28.

Table 4.26: Results of multivarient logistic regression model 5.

	β_0	**Std. Error**	**z-value**	**Signif.**
Intercept	35.6438	18.5317	1.923	0.05443
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2807	0.0760	3.694	0.0002
Gradient	-0.0405	0.0184	-2.208	0.0272
Months with flows ≥ 80th %ile (max. flows)	-0.1032	0.0359	-2.875	0.0040
Latitude	-0.6410	0.3353	-1.912	0.0559

Table 4.27: Summary statistics for model 5 (comparison with H_0).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
5	112.1	0.0000	0.2474	0.2903	0.3871	349.2	387.5	1.324

Table 4.28: Summary statistics for reduced model 5 (comparison with models 4a2 and 4b2).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
4a2	1.248	0.2640	0.0037	0.0038	0.0059	341.9	351.5	1.330
4b2	0.294	0.5875	0.0009	0.0009	0.0014	342.9	352.4	1.320

Predictive performance for model 5 (see Table 4.29) was worse than either of its precursor models (4a2 and 4b2).

Table 4.29: Predictive performance of model 5 compared with precursor models.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
4a2	66.05	78.18	74.83	70.11	72.17	0.4458
4b2	65.43	79.39	75.71	70.05	72.48	0.4529
5	64.81	79.39	75.54	69.68	72.17	0.4471
Testing data						
4a2	66.87	92.02	89.34	73.53	79.45	0.6085
4b2	67.48	92.64	90.16	74.02	80.06	0.6212
5	66.26	90.80	87.80	72.91	78.53	0.5885

All of the parameters remaining in model 5 were either extremely significant (e.g., for *Watershed area*, $\alpha = 0.0002$), or very close to the significance threshold (e.g., for *Latitude*, $\alpha = 0.0559$). As such, it was expected that model 5 was a sufficiently parsimonious model. However, in order to confirm that the remaining non-significant parameter (*Latitude*) was useful, a sixth round of model reduction was undertaken.

Sixth round model refinement: Model 6 eliminated the *Latitude* parameter from the model, such that the only remaining parameters were:

- Watershed area
- Gradient
- Months with flows $\geq$ the 80th percentile of maximum monthly flows

Table 4.30: Results of multivarient logistic regression model 6.

	β_0	Std. Error	z-value	Signif.
Intercept	0.2246	0.3277	0.685	0.4931
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.2754	0.0753	3.659	0.0003
Gradient	-0.0329	0.0177	-1.860	0.0629
Months with flows $\geq$ 80th %ile (max. flows)	-0.1108	0.0355	-3.125	0.0018

Table 4.31: Summary statistics for model 6 (comparison with H_0).

Model no.	$-2LL$	Signif.	R^2_L	R^2_{CS}	R^2_N	AIC	BIC	Odds Ratio
6	108.4	0.0000	0.2392	0.2822	0.3763	350.8	379.6	1.317

Table 4.32: Summary statistics for reduced model 6 (comparison with model 5).

Model no.	$-2LL$	Signif.	R^2_L	R^2_{CS}	R^2_N	AIC	BIC	Odds Ratio
5	3.687	0.0548	0.0107	0.0112	0.0172	343.2	352.7	1.324

Significance of the parameters under model 6 stayed similar to model 5, though the values for the *Gradient* parameter worsened. However, significance associated with the estimated intercept was worse under model 6, increasing from 0.0544 to 0.4931. R^2 coefficients were also worse under model 6, and although there had been consistent decreases in these values throughout model parameter eliminations, the drop was greater than usual for elimination of a single parameter. AIC under model 6 was worse, but BIC was better. Comparison of the reduced model 6 with model 5 confirmed that the *Gradient* parameter did contribute substantially to the model.

Table 4.33: Predictive performance of model 6 compared with model 5.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
5	64.81	79.39	75.54	69.68	72.17	0.4471
6	65.43	76.97	73.61	69.40	71.25	0.4270
Testing data						
5	66.26	90.80	87.80	72.91	78.53	0.5885
6	63.80	87.12	83.20	70.65	75.46	0.5236

While there was a slight increase in predictive success for fish-bearing streams in the modelling data set, all other measures were worse under model 6, compared with model 5.

Thus, elimination of *Gradient* from the model was deemed inadvisable, and model 5 considered a reasonably good, parsimonious model.

Re-checking for Multicollinearity

In Section 4.2.1, model 2 was checked for potential multicollinearity using VIF. While no specific parameter in model 2 exceeded thresholds of concern for VIF, the average VIF indicated that bias from multicollinearity might still be affecting the model. In order to check if the removal of a number of parameters has influenced these results for model 5, VIF was used to check this model. Once again, no specific parameter in the model exceeded thresholds of concern. However, although the average VIF had fallen from 1.73 in model 2 to 1.14 in model 5, this value is still above the threshold of 1, suggesting potential bias from multicollinearity.

4.2.2 Automated Parameter Refinement

Model refinement was undertaken manually, in order to consider a broad range of indications of model usefulness and therefore avoid relying on a single indicator to judge a "best" model—a failing that Whittingham et al. (2006) notes is common in ecological modelling. However, Burnham & Anderson (2002) suggest that AIC could defensibly be used as such a single indicator. The `step` command in *R* allows automated refinement of an input model, based in minimisation of AIC. As a check against the the manual approach undertaken, this method was used to refine model 2—the model refined by removal of highly correlated parameters in model 1.

Using the `step` command produced a "best" model with the same parameters as model 5, confirming that the results of manual model refinement were not inconsistent with refinement by AIC alone.

In addition to selecting parameters based on minimising AIC, `step` can also base selection on other indicators. For comparison, `step` was also run using selection based on minimisation of BIC. Automated refinement of model 2 using this method resulted in a model (model 7) with just two parameters:

- Watershed area
- Months with flows $\geq$ the 80th percentile of maximum monthly flows

Model 7 was significantly worse than model 5 on almost all measures other than BIC, and was not considered further.

Table 4.34: Results of multivarient logistic regression model 7 - automated selection by BIC.

	β_0	Std. Error	z-value	Signif.
Intercept	-0.0937	0.2813	-0.333	0.7390
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.3028	0.0759	3.990	6.61×10^{-5}
Months with flows $\geq$ 80th %ile (max. flows)	-0.1092	0.0353	-3.096	0.0020

Table 4.35: Summary statistics for model 7 - automated selection by BIC (comparison with H_0).

Model no.	$-2LL$	Signif.	R^2_L	R^2_{CS}	R^2_N	AIC	BIC	Odds Ratio
7	104.8	0.0000	0.2311	0.2741	0.3655	352.5	371.7	1.354

Table 4.36: Predictive performance of model 7 (automated selection by BIC) compared with model 5.

Model no.	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Modelling data						
5	64.81	79.39	75.54	69.68	72.17	0.4471
7	60.49	82.42	77.17	68.00	71.56	0.4403
Testing data						
5	66.26	90.80	87.80	72.91	78.53	0.5885
7	59.51	87.73	82.91	68.42	73.62	0.4924

4.3 Chapter Summary

Seven rounds of modelling and 13 individual models were tested to identify the most useful parameters for the model. The final parameter set—identified in model 5—included:

- Watershed area
- Gradient
- Months with flows ≥ the 80th percentile of maximum monthly flows
- Latitude

Model statistics for each model constructed are summarised in Table 4.37, while the measures of predictive performance for each model are summarised in Table 4.38.

Table 4.37: Summary statistics for all models (comparison with H_0).

Model no.	$-2LL$	**Signif.**	R^2_L	R^2_{CS}	R^2_N	**AIC**	**BIC**	**Odds Ratio**
1	146.7	0.000	0.3236	0.3615	0.4820	356.6	596.1	2.006
2	116.4	0.000	0.2568	0.2995	0.3994	350.9	417.9	1.210
3a	115.0	0.000	0.2537	0.2965	0.3953	350.3	407.8	1.255
3b	114.8	0.000	0.2532	0.2960	0.3947	350.5	408.0	1.214
4a1	113.1	0.000	0.2495	0.2924	0.3899	350.2	398.1	1.264
4a2	113.4	0.000	0.2501	0.2930	0.3907	349.9	397.8	1.330
4a3	113.0	0.000	0.2494	0.2923	0.3897	350.2	398.1	1.225
4a4	112.3	0.000	0.2477	0.2906	0.3875	351.0	398.9	1.260
4b2	112.4	0.000	0.2480	0.2909	0.3879	350.9	398.8	1.320
4b3	111.4	0.000	0.2459	0.2888	0.3851	351.8	399.7	1.200
5	112.1	0.0000	0.2474	0.2903	0.3871	349.2	387.5	1.324
6	108.4	0.0000	0.2392	0.2822	0.3763	350.8	379.6	1.317
7	104.8	0.0000	0.2311	0.2741	0.3655	352.5	371.7	1.354

Table 4.38: Predictive performance of all models.

Model no.	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**	**MCC**
Modelling data						
1	64.81	87.27	83.33	71.64	76.15	0.5351
2	70.37	77.58	75.50	72.73	74.01	0.4808
3a	67.90	78.18	75.34	71.27	73.09	0.4635
3b	69.75	78.18	75.84	72.47	74.01	0.4812
4a1	67.28	80.00	76.76	71.35	73.70	0.4770
4a2	66.05	78.18	74.83	70.11	72.17	0.4458
4a3	64.20	76.97	73.24	68.65	70.64	0.4153
4a4	62.35	79.39	74.81	68.23	70.95	0.4239
4b2	65.43	79.39	75.71	70.05	72.48	0.4529
4b3	66.05	78.79	75.35	70.27	72.48	0.4523
5	64.81	79.39	75.54	69.68	72.17	0.4471
6	65.43	76.97	73.61	69.40	71.25	0.4270
7	60.49	82.42	77.17	68.00	71.56	0.4403
Testing data						
1	58.28	86.50	81.20	67.46	72.39	0.4668
2	70.55	87.12	84.56	74.74	78.83	0.5848
3a	69.33	87.73	84.96	74.09	78.53	0.5805
3b	67.48	87.12	83.97	72.82	77.30	0.5568
4a1	66.26	87.12	83.72	72.08	76.69	0.5457
4a2	66.87	92.02	89.34	73.53	79.45	0.6085
4a3	63.80	87.73	83.87	70.79	75.77	0.5308
4a4	64.42	88.34	84.68	71.29	76.38	0.5434
4b2	67.48	92.64	90.16	74.02	80.06	0.6212
4b3	65.03	85.89	82.17	71.07	75.46	0.5207
5	66.26	90.80	87.80	72.91	78.53	0.5885
6	63.80	87.12	83.20	70.65	75.46	0.5236
7	59.51	87.73	82.91	68.42	73.62	0.4924

The coefficients calculated for model 5 (see Table 4.26) yield a model of the form:

$$\pi_{fish\text{-}bearing} = \frac{e^{35.6438+0.2807A-0.0405G-0.1032F_{80max}-0.6410L}}{1+e^{35.6438+0.2807A-0.0405G-0.1032F_{80max}-0.6410L}} \tag{4.4}$$

where:

$\pi_{fish\text{-}bearing}$ = Probability of being fish-bearing

A = Watershed area

G = Gradient

F_{80max} = Months with flows $\geq$ the 80th percentile of maximum monthly flows

L = Latitude

Chapter 5

Model Validation

5.1 Validation with the Test Data Set

Simple validation of models was performed throughout the parameter elimination phase of model refinement. Each time a new model was generated, its predictive performance was checked on the partitioned testing data set—an approach referred to as the *validation set approach* (James, Witten, Hastie & Tibshirani 2013). For most models, this testing demonstrated some variability in model performance between the modelling data and the testing data. For model 5, these differences are shown through the confusion matrices for model 5 results when run against the modelling data set (see Table 5.1) and when run against the testing data set (see Table 5.2).

Table 5.1: Confusion matrix for model 5 results run against model data set.

	Non-fish-bearing	Fish-bearing	Total
Predicted non-fish-bearing	131	57	188
Predicted fish-bearing	34	105	139
Total	188	139	327

From the confusion matrices, metrics for predictive performance could be produced. These were generated for each model throughout model development (see Section 4.3). Measures of predictive performance for model 5 are reproduced in Table 5.3.

Murphy & Winkler (1987), as cited in Pearce & Ferrier (2000), note that predictive

Table 5.2: Confusion matrix for model 5 results run against testing data set.

	Non-fish-bearing	Fish-bearing	Total
Predicted non-fish-bearing	148	55	203
Predicted fish-bearing	15	108	123
Total	163	163	326

Table 5.3: Predictive performance of model 5.

Model no.	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**	**MCC**
Modelling data						
5	64.81	79.39	75.54	69.68	72.17	0.4471
Testing data						
5	66.26	90.80	87.80	72.91	78.53	0.5885

performance can be visualised by inspecting the overlap of distribution of both binary responses plotted on the same axis, as shown in Figure 5.1. This plot shows four distribution curves: two for each data set. The two curves on the left-hand side of the plot relate to those sites within each data set that have been classified as non-fish-bearing. The curves show the occurrence frequency for the probabilities predicted by the model. As would be expected, most of the predicted probabilities for the non-fish-bearing streams are below 0.5—that is, the model predicts that for most of these streams, there is a less than 50% chance that these streams are fish-bearing. Probabilities peak at around 0.2, with a smaller peak around 0.45. The tail of the curves does extend above 0.5, which accounts for the 10% to 20% of non-fish-bearing streams not correctly predicted by the model.

Similarly, the two curves on the right-hand side of the plot relate to those sites within each data set that have been classified as fish-bearing. These curves show the occurrence frequency for the probabilities predicted by the model for these fish-bearing sites. For these curves, most of the probabilities are above 0.5, though less so than for the non-fish-bearing streams. Peaks occur at around 0.9 and around 0.5. This means that for a substantial proportion of the fish-bearing sites, the model produces a probability of less than 0.5 that the streams are fish-bearing—that is, it incorrectly predicts that these sites are non-fish-bearing streams.

The peaks in distributions in the predicted probability range of 0.4 to 0.6 indicate that the model was relatively uncertain for a substantial proportion of the sites. The shape of distributions for results for both data sets were similar, indicating that each data set was reasonably representative of the combined data set. This was reinforced by the similarities in distributions between the data sets for each of the models produced throughout model development (see Appendix E). For model 5, the probability distributions for fish-bearing sites were very similar. For non-fish-bearing sites, though the distribution had a similar shape, there was greater variation in the peak height and spread of the distribution curve. This was consistent with the results for predictive performance, which showed greater variation for non-fish-bearing streams—specificity varied from 79.39% to 90.80%—than it did for fish-bearing streams—sensitivity varied from 64.81% to 66.26%.

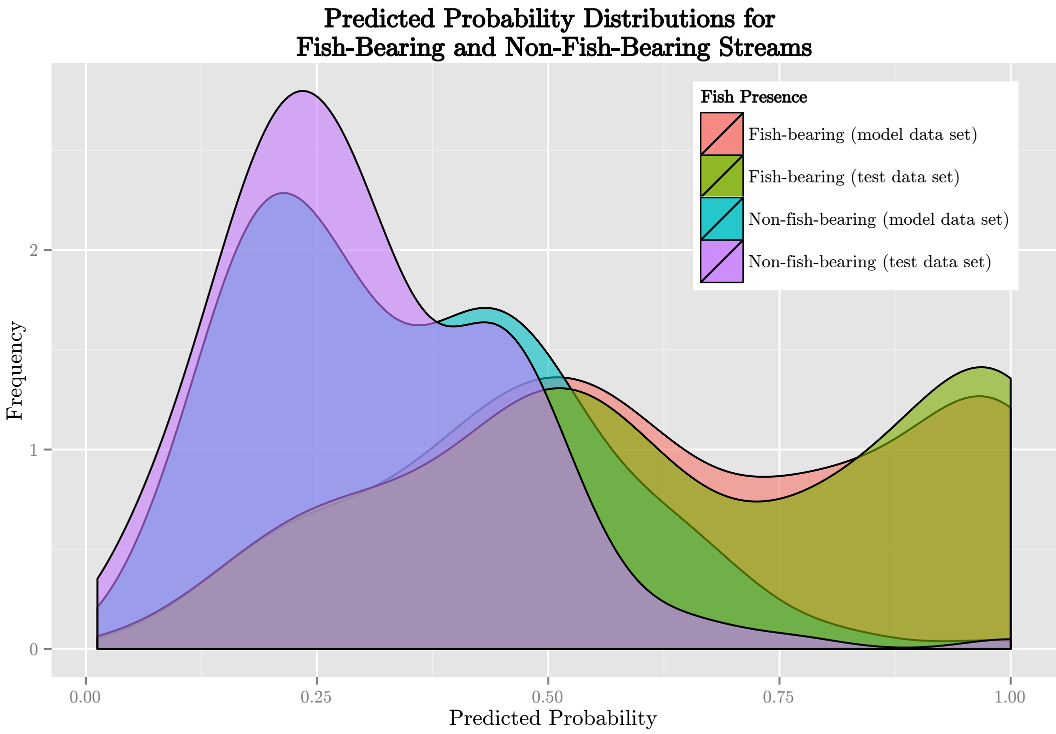


Figure 5.1: Overlapping distributions of probabilities frequencies from model 5 for both non-fish-bearing and fish-bearing streams (model and testing data sets).

5.2 Model Cross-Validation

For each of the preliminary models, validation by checking model results against the testing data set was used purely to assist in parameter selection, so no more exhaustive efforts at cross-validation were undertaken. However, once the preferred modelling parameter set was identified (model 5), additional cross-validation of the model was conducted.

5.2.1 Remodelling for Cross-Validation

To prepare for cross-validation, the partitioned data sets (modelling and testing) were combined into one data set. A new model, 5c, was generated using those parameters identified in model 5, but fitted against the combined data set (modelling data set plus testing data set). Results from model 5c are summarised in Table 5.4. Table 5.5 provides a comparison of the model coefficients for model 5 and model 5c. Comparison of model 5c with the null hypothesis yielded the statistics summarised in Table 5.6. Model statistics for the original model 5 are also included in this table for comparison.

Table 5.4: Results of multivarient logistic regression model 5c.

	β_0	**Std. Error**	**z-value**	**Signif.**
Intercept	64.2415	13.6934	4.691	2.71×10^{-6}
Input Parameter	β_1	**Std. Error**	**z-value**	**Signif.**
Watershed area	0.1757	0.0420	4.182	2.89×10^{-5}
Gradient	-0.0637	0.0129	-4.930	8.24×10^{-7}
Months with flows $\geq$ 80th %ile (max. flows)	-0.1255	0.0250	-5.017	5.25×10^{-7}
Latitude	-1.1491	0.2475	-4.643	3.43×10^{-6}

Model 5c was also tested for predictive performance against the combined data set. However, as this same data set was used to train the model, outcomes may have overstated the effectiveness of the model. The confusion matrix in Table 5.7 shows the outcomes of the model. Table 5.8 summarises predictive success measures. Results for model 5 are also included for comparison. Matthews correlation coefficient was not able to be calculated for the model outcomes run against the combined data set. The probability distribution for model 5c is shown in Figure 5.2. For comparison, the probability distribution of model 5 against the combined data set is shown in Figure

5.3.

Table 5.5: Comparison of model coefficients - model 5 and model 5c.

	$\boldsymbol{\beta_0(5)}$	$\boldsymbol{\beta_0}$ (5c)
Intercept	35.6438	64.2415
Input Parameter	$\boldsymbol{\beta_1(5)}$	$\boldsymbol{\beta_1(5c)}$
Watershed area	0.2807	0.1757
Gradient	-0.0405	-0.0637
Months with flows $\geq$ 80th %ile (max. flows)	-0.1032	-0.1255
Latitude	-0.6410	-1.1491

Table 5.6: Summary statistics for model 5c (comparison with H_0).

Model no.	**Likelihood ratio**	**Signif.**	$\boldsymbol{R^2_L}$	$\boldsymbol{R^2_{CS}}$	$\boldsymbol{R^2_N}$	**AIC**	**BIC**	**Odds Ratio**
5	112.1	0.0000	0.2474	0.2903	0.3871	349.2	387.5	1.324
5c	241.8	0.0000	0.2671	0.3095	0.4126	671.4	715.3	1.192

Table 5.7: Confusion matrix for model 5c results run against the combined data set.

	Non-fish-bearing	Fish-bearing	Total
Predicted non-fish-bearing	260	95	355
Predicted fish-bearing	68	230	298
Total	355	298	653

Table 5.8: Predictive performance of model 5c compared with model 5.

Model (Data set)	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**	**MCC**
5c (combined)	70.77	79.27	77.18	73.24	75.04	NA
5 (combined)	65.54	85.06	81.30	71.36	75.34	NA
5 (modelling)	64.81	79.39	75.54	69.68	72.17	0.4471
5 (testing)	66.26	90.80	87.80	72.91	78.53	0.5885

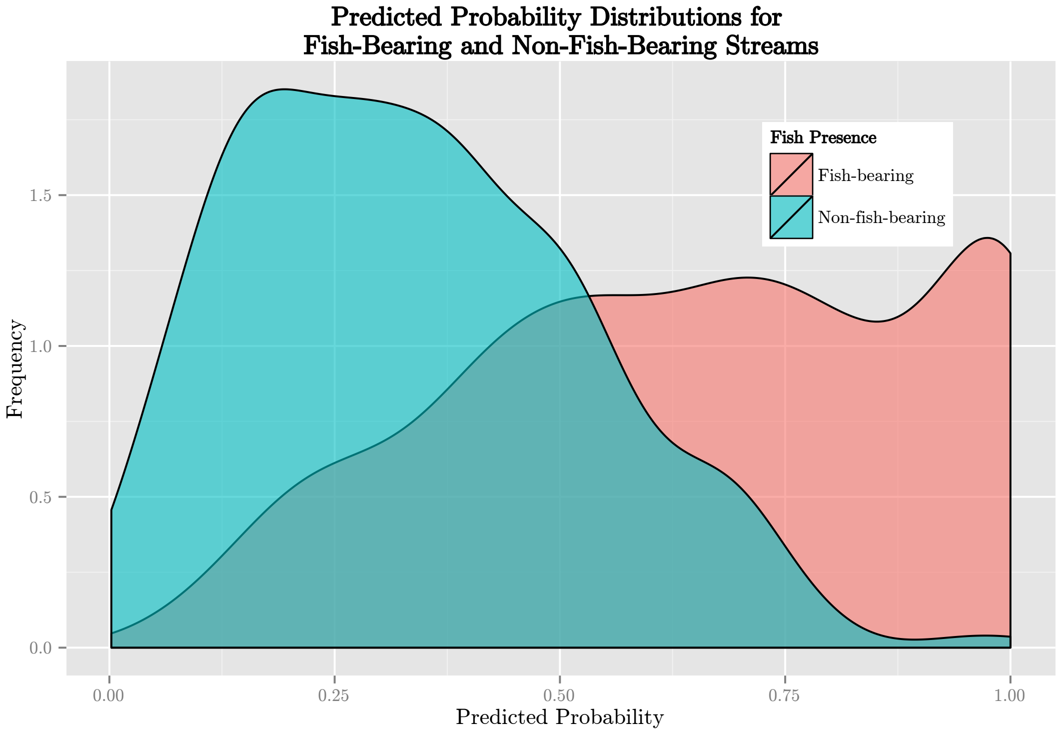


Figure 5.2: Overlapping distributions of probability frequencies from model 5c (combined data set).

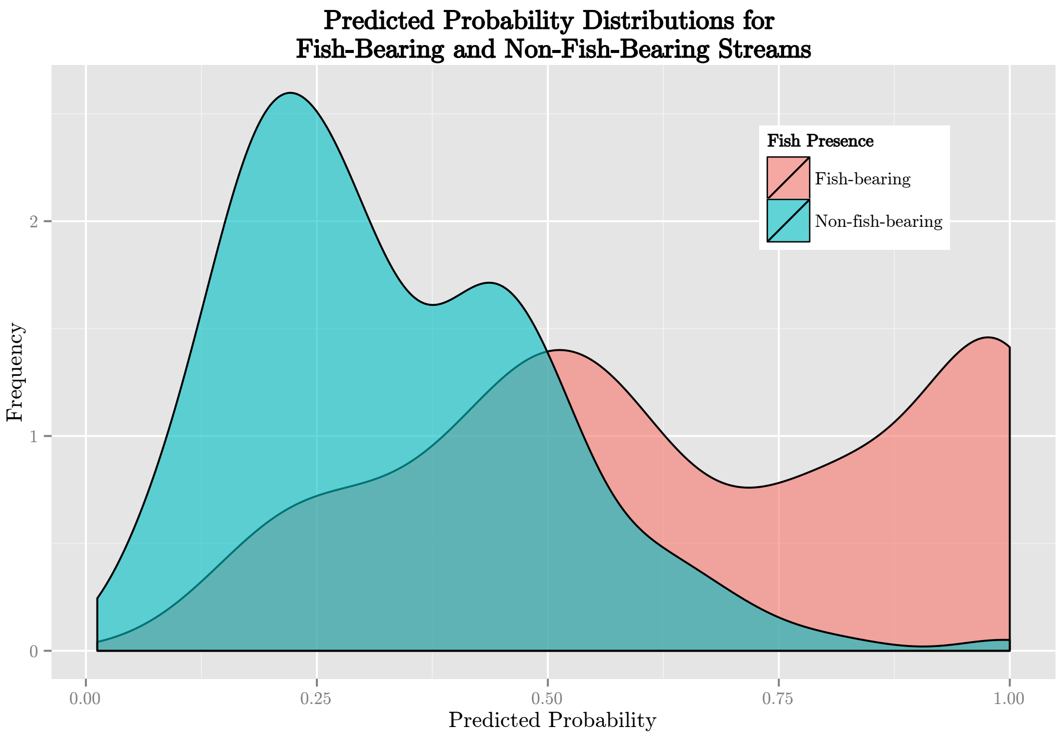


Figure 5.3: Overlapping distributions of probability frequencies from model 5 (combined data set).

5.2.2 Cross-Validation Methods

As the combined data set was still not overly large (653 sites), *leave-one-out* cross-validation (LOOCV) was employed. LOOCV eliminates potential variation from randomness in the selection of the split in sets, and eliminates overestimation of error rates that can be produced by the validation approach (James et al. 2013). James et al. (2013) also suggests that k-fold cross-validation with $k = 5$ and $k = 10$ can have more accurate results for test error than LOOCV, so 5-fold and 10-fold cross-validation was also undertaken for comparison.

Cross-validation was performed using the `cv.glm` command in R. The `cost` function used within `cv.glm` was taken from Weiss (2009):

```
cost <- function(r, pi=0) mean(abs(r-pi)>0.5)
```

The `cv.glm` command in R produces a delta value and adjusted delta value, where the delta value is cross-validation misclassification error (Weiss 2009) and the adjusted delta value modifies the delta value to account for bias produced by using k-fold cross-validation rather than LOOCV (James et al. 2013, Weiss 2009). Delta values and adjusted delta values for each approach to cross-validation are compared in Table 5.9.

Table 5.9: Results of cross-validation of model 5c.

	LOOCV	**5-fold**	**10-fold**
Delta	0.2527	0.2481	0.2588
Adjusted delta	0.2526	0.2444	0.2563

The delta values give a measure of model error. Adjusted delta values from the various cross-validation methods were all reasonably close, ranging from 0.2444 to 0.2563. This indicates an error level in the predicted probability outcomes of model 5c of around 25%. This error level is reasonably high, but not unexpected given that the overall accuracy of model 5c when tested against the combined data set was 75.04%.

5.3 Comparison of Model 5 and Model 5c

Unlike model 5, model 5c used all available data to refine parameter coefficient values. It was therefore expected to perform better than model 5 in terms of predictive success. This was not observed (see Table 5.8). While model 5c had substantially better success in correctly identifying fish-bearing streams (by around 5%), this was almost exactly offset by a similar reduction in correctly predicting non-fish-bearing streams. Overall accuracy of the models was virtually identical.

Given that accurate prediction of non-fish-bearing streams is more useful for field planning than prediction of fish-bearing streams, model 5 is preferred over model 5c. Accurate identification of non-fish-bearing streams by the model would allow prioritisation of those streams for field surveys in order to dedicate survey resources to meeting regulatory requirements for assigning non-fish-bearing status to those streams, and not "wasting" resources on those streams less likely to be non-fish-bearing.

Chapter 6

Results and Conclusions

6.1 Model Usefulness

Predictive success rates for model 5 were not exceptional, but were sufficiently high for the model to be a useful tool for field planning. Predictive success rates were consistently higher for non-fish-bearing streams than for fish-bearing streams, including in the final model. This difference was likely because of underlying differences in the quality of the fish-bearing classification data. As noted in Section 2, streams classified as non-fish-bearing during assessments for environmental assessments must meet very stringent guidelines.

If the conditions for classification as a non-fish-bearing stream are not met, then classification of the stream defaults to fish-bearing, irrespective of evidence to the contrary. Thus, sites classified as non-fish-bearing have far higher certainty in their classification, and therefore lower error than those sites classified as fish-bearing. That is, very few sites classified as non-fish-bearing are likely to actually be fish-bearing, but a much higher proportion of sites classified as fish-bearing may actually be non-fish-bearing.

The effects of this higher degree of error in the sites classified as fish-bearing was demonstrated by the distribution predicted probabilities for all models (see Appendix E). For all models generated throughout model development, the distribution of probability for the fish-bearing sites was flatter and wider than the distributions for the non-fish-bearing sites.

Although the lower predictive success of the model for fish-bearing streams reduces its usefulness, accurate prediction of non-fish-bearing streams is more useful for field planning than prediction of fish-bearing streams. Given that predictive success for non-fish-bearing streams would be a priority for field planning, the model outputs could be further biased to increase predictive success for these sites, while sacrificing predictive success for fish-bearing sites. This could be simply implemented by shifting the threshold of prediction higher from the threshold (probability = 0.5) used by default in the model. Hoever, the disadvantage of shifting the threshold higher would be an increase in "false negatives", that is, fish-bearing streams incorrectly identified as non-fish-bearing. Given the peak in the fish-bearing probability distribution for model 5 near 0.5 (see Figure 5.3), the value in threshold shifting is questionable.

6.2 Further Work

As noted in Section 3, one of the parameters in the final model (*gradient*) was not desktop-available data, but was used as a proxy for desktop data derived slope information for higher resolution elevation data not available for this research. In order to confirm the validity of the model as a purely desktop analysis, the model usefulness should be confirmed using actual desktop-available slope data.

Additionally, a number of other potential predictive parameters could also be investigated. The shape of distribution curves for many of the models (see Appendix E) seemed to indicate that substantial variation is not accounted for by the selected parameters. Discussions with professional colleagues (Mitchell, S 2014, pers. comm., 5 September) has suggested that longer-term, intermittent, inter-annual hydrologic events such as recurring droughts or floods may strongly influence fish-bearing status. Data on these types of events is obscured in the data used for model development to date, by the averaging used to calculate mean and maximum monthly flows.

Basic climate data such as rainfall, snowfall and temperature may also influence fish-bearing status and could be included in future analysis.

6.3 Conclusion

With predictive success rates for non-fish-bearing streams (specificity) in the range of approximately 80% to 90%, model 5 could be very useful for field planning purposes. However, before it could be utilised in the manner envisaged at the beginning of this research—that is, based purely on desktop-available data—model performance needs to be confirmed using higher-quality slope data, derived from finer-grained elevation data.

Poorer predictive success rates for fish-bearing streams (sensitivity) limit the model's usefulness for other purposes. While this may be addressed by identifying and including other parameters in the model (see Section 6.2), accuracy is likely unavoidably handicapped by biases in data quality caused by the regulatory regime under which stream classification occurs.

Overall, the model could be a very useful tool but should be further validated and refined before serious implementation.

References

BC Fisheries Information Services Branch (2000), *Reconnaissance (1:20,000) Fish and Fish Habitat Inventory: Users Guide to the Fish and Fish Habitat Assessment Tool (FHAT20)*, 1.0 edn, Resources Inventory Committee, BC.
URL: *http://www.for.gov.bc.ca/hts/risc/pubs/aquatic/fhat20/assets/fhat20.pdf*

BC Fisheries Information Services Branch (2001), *Reconnaissance (1:20,000) Fish and Fish Habitat Inventory: Standards and Procedures*, 2.0 edn, Resources Inventory Committee, BC.
URL: *http://www.for.gov.bc.ca/hts/risc/pubs/aquatic/recon/recce2c.pdf*

BC Ministry of Environment (n.d.), 'Fisheries inventory - field data information system (FDIS) introduction'.
URL: *http://www.env.gov.bc.ca/fish/fdis/description.html*

BC Oil and Gas Commission (2013), Environmental protection and management guide, Technical report.
URL: *http://www.bcogc.ca/node/5899/download*

Beanlands, G. E. & Duinker, P. N. (1983), An ecological framework for environmental impact assessment in canada, Technical report, Institute for Resource and Environmental Studies, Dalhousie University and Federal Environmental Assessment Review Office.
URL: *http://epe.lac-bac.gc.ca/100/200/301/ceaa-acee/ecological*$_f$*ramework* $-$ *e/*23*E.PDF*

Bowerman, B. & O'Connell, R. (1990), *Linear statistical models: An applied approach*, 2nd edn, Duxbury, Belmont, California.

Burnham, K. P. & Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn, Springer, New York, NY.

Canadian Environmental Assessment Agency (2009), 'Canadian environmental assessment agency - policy and guidance - glossary'.

URL: *http://www.ceaa-acee.gc.ca/default.asp?lang=Enn=B7CA7139-1offset=3toc=hidevec*

Dalgaard, P. (2009), *Introductory Statistics with R*, 2 edn, Springer, [S.l.].

Environmental Assessment Office (2013), Guideline for the selection of valued components and assessment of potential effects, Technical report.

URL: *http://www.eao.gov.bc.ca/pdf/EAO*$_V$*alued*$_C$*omponents*$_G$*uideline*$_2$013$_0$9$_0$9*.pdf*

Field, A. P. (2012), *Discovering statistics using R*, Sage, London ; Thousand Oaks, Calif.

Filipe, A. F., Cowx, I. G. & Collares-Pereira, M. J. (2002), 'Spatial modelling of freshwater fish in semi-arid river systems: a tool for conservation', *River Research and Applications* **18**(2), 123–136.

URL: *http://doi.wiley.com/10.1002/rra.638*

Fisheries and Oceans Canada (2013), 'Fisheries protection policy statement'.

URL: *http://www.dfo-mpo.gc.ca/pnw-ppe/pol/index-eng.html*

Forest Service British Columbia (1998), *Fish-stream Identification Guidebook (Forest Practices Code of British Columbia)*, 2.1 edn.

URL: *https://www.for.gov.bc.ca/tasb/legsregs/fpc/fpcguide/FISH/FishStream.pdf*

Gotelli, N. J. (2004), *A primer of ecological statistics*, Sinauer Associates Publishers, Sunderland, Mass.

Government of Canada (2013*a*), 'Canadian environmental assessment act, 2012'.

URL: *http://laws-lois.justice.gc.ca/eng/acts/c-15.21/FullText.html*

Government of Canada (2013*b*), 'Fisheries act'.

URL: *http://laws-lois.justice.gc.ca/eng/acts/f-14/FullText.html*

Government of Canada (2013*c*), 'Species at risk act'.

URL: *http://laws-lois.justice.gc.ca/eng/acts/s-15.3/FullText.html*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Vol. 103 of *Springer Texts in Statistics*, Springer New York, New York, NY.

URL: *http://link.springer.com/10.1007/978-1-4614-7138-7*

Joy, M. K. & Death, R. G. (2000), 'Development and application of a predictive model of riverine fish community assemblages in the taranaki region of the north island, new zealand', *New Zealand Journal of Marine and Freshwater Research* **34**(2), 241–252.
URL: *http://www.tandfonline.com/doi/abs/10.1080/00288330.2000.9516930*

Joy, M. K. & Death, R. G. (2002), 'Predictive modelling of freshwater fish as a biomonitoring tool in new zealand', *Freshwater Biology* **47**(11), 2261–2275.
URL: *http://doi.wiley.com/10.1046/j.1365-2427.2002.00954.x*

Joy, M. K. & Death, R. G. (2004), 'Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks', *Freshwater Biology* **49**(8), 1036–1052.
URL: *http://doi.wiley.com/10.1111/j.1365-2427.2004.01248.x*

Mastrorillo, S., Lek, S., Dauba, F. & Belaud, A. (1997), 'The use of artificial neural networks to predict the presence of small-bodied fish in a river', *Freshwater Biology* **38**(2), 237–246.
URL: *http://doi.wiley.com/10.1046/j.1365-2427.1997.00209.x*

Mugodo, J., Kennard, M., Liston, P., Nichols, S., Linke, S., Norris, R. H. & Lintermans, M. (2006), 'Local stream habitat variables predicted from catchment scale characteristics are useful for predicting fish distribution', *Hydrobiologia* **572**(1), 59–70.
URL: *http://link.springer.com/10.1007/s10750-006-0252-7*

Murphy, A. H. & Winkler, R. L. (1987), 'A general framework for forecast verification', *Monthly Weather Review* **115**, 1330–1338.

Myers, R. (1990), *Classical and modern regression with applications*, 2nd edn, Duxbury, Boston.

Olden, J. D. & Jackson, D. A. (2002), 'A comparison of statistical approaches for modelling fish species distributions', *Freshwater Biology* **47**(10), 1976–1995.
URL: *http://doi.wiley.com/10.1046/j.1365-2427.2002.00945.x*

Pearce, J. & Ferrier, S. (2000), 'Evaluating the predictive performance of habitat models developed using logistic regression', *Ecological modelling* **133**(3), 225–245.
URL: *http://www.sciencedirect.com/science/article/pii/S0304380000003227*

Porter, M. S., Rosenfeld, J. & Parkinson, E. A. (2000), 'Predictive models of fish species distribution in the blackwater drainage, british columbia', *North American Journal of Fisheries Management* **20**(2), 349–359.

Province of British Columbia (2008), 'Oil and gas activities act'.
URL: *http://www.bclaws.ca/civix/document/id/complete/statreg/08036*$_0$1

Province of British Columbia (2013), 'Environmental protection and management regulation'.
URL: *http://www.bclaws.ca/civix/document/id/complete/statreg/200*$_2$010

Quinn, G. P. (2002), *Experimental design and data analysis for biologists*, Cambridge University Press, Cambridge, UK ; New York.

Stantec Consulting Ltd. (2014), Prince rupert gas transmission project - application for an environmental assessment certificate, Technical Report PRGT004776 - TC - EN - FM - 0001.
URL: *http://a100.gov.bc.ca/appsdata/epic/html/deploy/epic*$_d$*ocument*$_4$03$_3$7577*.html*

Walters, C. & Ludwig, D. (1994), 'Calculation of bayes posterior probability distributions for key population parameters', Canadian Journal of Fisheries and Aquatic Science *(51), 713–722.*
URL: *http://www.nrcresearchpress.com/doi/pdf/10.1139/f94-071*

Weiss, J. (2009), 'Lecture 19Monday, October 27, 2008'.
URL: *http://www.unc.edu/courses/2008fall/ecol/563/001/docs/lectures/lecture19.htm*

Whitlock, M. (2009), The analysis of biological data*, Roberts and Co. Publishers, Greenwood Village, Colo.*

Whittingham, M. J., Stephens, P. A., Bradbury, R. B. & Freckleton, R. P. (2006), 'Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour', Journal of Animal Ecology **75***(5), 1182–1189.*
URL: *http://doi.wiley.com/10.1111/j.1365-2656.2006.01141.x*

Zar, J. H. (1996), Biostatistical analysis*, 3rd edn, Prentice Hall, Upper Saddle River, N.J.*

Appendix A

Project Specification

University of Southern Queensland

FACULTY OF ENGINEERING AND SURVEYING

ENG4111/4112 Research Project
PROJECT SPECIFICATION
Version 2
30 March 2014

FOR: **Benjamin James BYRD**

TOPIC: Using desktop hydrologic data to predict fish presence in streams in northern British Columbia

SUPERVISORS: Dr. Ian Brodie, USQ
Heidi Biberhofer, P.Eng., MRM, Senior Water Resources Engineer, Stantec Consulting Ltd.
Kirby Ottenbriet, B.A. Fisheries Technical Lead, Stantec Consulting Ltd.

SPONSORSHIP: Stantec Consulting Ltd.

CONFIDENTIALITY: Proprietary Stantec and client data can be used for thesis production. Other publication would require approval of all applicable parties.

PROJECT AIMS: To create a method for using desktop hydrologic data collected and analysed during environmental assessments to predict fish presence in streams in British Columbia, for more efficient allocation of ground-truthing field work by fisheries biologists.

PROGRAMME:

1. Research the follow elements related to the project:
   - Regulatory requirements for stream classification in British Columbia
   - Biological approach to classifying streams of “fish bearing” or “not fish bearing”
   - Hydrologic parameters considered important to fish biologists in determining stream status
2. Obtain, collate and organize existing hydrologic and stream classification data for stream crossing analysed as part of one or more pipeline environmental assessment projects in British Columbia.
3. Investigate collated data sets for correlations between hyrdologic data and stream classification, guiding by initial research results.
4. Analysis correlations between data sets, and identify if additional data might be needed for predictive modelling.
5. Develop a method to predict if a steam is likely fish-bearing or likely not-fish-bearing.
6. Compare method results with actual field stream classification data for initial validation.
7. Test predictive model on an independent data set for secondary validation.

As time permits:

8. Produce GIS mapping indicating predicted fish presence and actual fish presence for both initial and independent data sets.

AGREED:

________________ (Student) ________________ , ________________ , ________________ (Supervisors)

/ / 2014 / / 2014 / / 2014 / / 2014

Appendix B

R Functions

B.1 R Function for single-parameter logistic regression

```
#Regress analysis set
logreg <- function (dataset, fishcol, datacol) {
  setname <- paste(deparse(substitute(dataset)))
  colname <- paste(deparse(substitute(datacol)))
  dataname <- paste(setname,colname,sep="-")
  filename <- paste(setname,"csv",sep=".")
  dataset <- read.csv(filename)
  attach(dataset)
  log_reg <- glm(fishcol ~ datacol,family=binomial,data=dataset)
  print(summary(log_reg))
  cat("Number of samples =", nrow(dataset),"\n")
  dev_base <- log_reg$null.deviance
  dev_new <- log_reg$deviance
  log_reg_chi <- dev_base - dev_new
  cat("Likeihood ratio =", log_reg_chi ,"\n")
  log_reg_chif <- log_reg$df.null - log_reg$df.residual
  log_reg_p <- 1 - pchisq(log_reg_chi, log_reg_chif)
  cat("Significance of LR =", log_reg_p ,"\n")
  r2l <- log_reg_chi/dev_base
  cat("R2L =", r2l ,"\n")
  r2cs <- 1 - exp((dev_new - dev_base)/nrow(dataset))
  cat("R2CS =", r2cs ,"\n")
  r2n <- r2cs/(1-exp(-(dev_base)/nrow(dataset)))
  cat("R2N =", r2n ,"\n")
  lr_x <- sort(datacol)
  lr_B0 <- coefficients(log_reg)[c(1)]
  lr_B1 <- coefficients(log_reg)[c(2)]
  odd_rat <- exp(lr_B1)
  cat("Odds ratio =", odd_rat,"\n")
  lr_pi <- lr_B0 + lr_B1*lr_x
  lr_y <- exp(lr_pi)/(1+exp(lr_pi))
  plot(fishcol~datacol)
  lines(lr_x,lr_y,col="red")
  newline <- data.frame(dataname,lr_B0,lr_B1,log_reg_chi,
          log_reg_p,r2l,r2cs,r2n,odd_rat)
  write.table(newline,file="IndStats.csv",sep=",",
          col.names=FALSE,append=TRUE)
  detach(dataset)
}
```

B.2 R Function for key model statistics (H_0)

```
#Compare logistic regression model against null hypothesis
model_test <- function (model) {
  specify_decimal <- function(x, k) format(round(x, k), nsmall=k)
  cat("Number of samples =", nobs(model),"\n")
  dev_base <- model$null.deviance
  dev_new <- model$deviance
  model_chi <- dev_base - dev_new
  cat("Likeihood ratio =", model_chi ,"\n")
  model_chif <- model$df.null - model$df.residual
  model_p <- 1 - pchisq(model_chi, model_chif)
  cat("Significance of LR =", model_p ,"\n")
  r2l <- model_chi/dev_base
  cat("R2L =", r2l ,"\n")
  r2cs <- 1 - exp((dev_new - dev_base)/nobs(model))
  cat("R2CS =", r2cs ,"\n")
  r2n <- r2cs/(1-exp(-(dev_base)/nobs(model)))
  cat("R2N =", r2n ,"\n")
  lr_B0 <- coefficients(model)[c(1)]
  lr_B1 <- coefficients(model)[c(2)]
  odd_rat <- exp(lr_B1)
  cat("Odds ratio =", odd_rat,"\n")
  Akaike_IC <- dev_new + 2*model_chif
  cat("Akaike information criterion =", Akaike_IC,"\n")
  Bayes_IC <- dev_new + 2*model_chif*(log(nobs(model)))
  cat("Bayes information criterion =", Bayes_IC,"\n")
  cat("LaTEX:  &",specify_decimal(model_chi,1) ,"&",
          specify_decimal(model_p,4) ,"&", specify_decimal(r2l,4),
          "&",specify_decimal(r2cs,4),"&",specify_decimal(r2n,4),
        "&",specify_decimal(Akaike_IC,1),"&",
        specify_decimal(Bayes_IC,1),"&",
        specify_decimal(odd_rat,3),"\n")
 }
```

B.3 R Function for key model statistics (reduced model)

```
#Compare logistic regression models; model1 has more variables than model2
model_comp <- function (model1,model2) {
  specify_decimal <- function(x, k) format(round(x, k), nsmall=k)
  cat("Number of samples =", nobs(model1),"\n")
  dev_base <- model2$deviance
  dev_new <- model1$deviance
  model1_chi <- dev_base - dev_new
  cat("Likeihood ratio =", model1_chi ,"\n")
  model1_chif <- model2$df.residual - model1$df.residual
  model1_p <- 1 - pchisq(model1_chi, model1_chif)
  cat("Significance of LR =", model1_p ,"\n")
  r2l <- model1_chi/dev_base
  cat("R2L =", r2l ,"\n")
  r2cs <- 1 - exp((dev_new - dev_base)/nobs(model1))
  cat("R2CS =", r2cs ,"\n")
  r2n <- r2cs/(1-exp(-(dev_base)/nobs(model1)))
  cat("R2N =", r2n ,"\n")
  lr_B0 <- coefficients(model1)[c(1)]
  lr_B1 <- coefficients(model1)[c(2)]
  odd_rat <- exp(lr_B1)
  cat("Odds ratio =", odd_rat,"\n")
  Akaike_IC <- dev_new + 2*model1_chif
  cat("Akaike information criterion =", Akaike_IC,"\n")
  Bayes_IC <- dev_new + 2*model1_chif*(log(nobs(model1)))
  cat("Bayes information criterion =", Bayes_IC,"\n")
  cat("LaTEX:  &",specify_decimal(model1_chi,3) ,"&",
  specify_decimal(model1_p,4) ,"&",
  specify_decimal(r2l,4),"&",specify_decimal(r2cs,4),
  "&",specify_decimal(r2n,4),"&", specify_decimal(Akaike_IC,1),
  "&",specify_decimal(Bayes_IC,1),
  "&",specify_decimal(odd_rat,3),"\n")
}
```

B.4 R Function for predictive performance analysis

```
#Check precentage true/false fish presence correctly predicted by model
pred_check2 <- function (model,data) {
  specify_decimal <- function(x, k) format(round(x, k), nsmall=k)
  fit_vals <- predict(model,newdata=data,type='response')
  real_vals <- data$fish + 0
  pred_vals <- round(fit_vals)
  cor_vals <- pred_vals + real_vals
  real_0 <- table(real_vals)[["0"]]
  real_1 <- table(real_vals)[["1"]]
  pred_0 <- table(pred_vals)[["0"]]
  pred_1 <- table(pred_vals)[["1"]]
  cor_0 <- table(cor_vals)[["0"]]
  cor_1 <- table(cor_vals)[["2"]]
  incor_0 <- real_0 - cor_0
  incor_1 <- real_1 - cor_1
  val_comp <- cbind(real_vals,fit_vals, deparse.level = 1)
  comp_0 <- subset(val_comp, real_vals == 0)
  comp_1 <- subset(val_comp, real_vals == 1)
  sens <- cor_1/real_1*100
  spec <- cor_0/real_0*100
  corAll <- (cor_1+cor_0)/(real_0+real_1)*100
  ppv <- cor_1/pred_1*100
  npv <- cor_0/pred_0*100
  mcc <- ((cor_1 * cor_0) - (incor_1 * incor_0)) /
          sqrt((cor_1 + incor_1)*(cor_1 + incor_0)*
          (cor_0 + incor_1)*(cor_0 + incor_0))
  cat("Sensitivity (Percentage fish-bearing correctly predicted)
          =", sens ,"\n")
  cat("Specificity (Percentage non-fish-bearing correctly predicted)
           =", spec ,"\n")
  cat("PPV (Percentage fish-bearing predictions correct)
          =", ppv ,"\n")
  cat("NPV (Percentage non-fish-bearing predictions correct)
           =", npv ,"\n")
  cat("Overall percentage correctly predicted =", corAll ,"\n")
  cat("Mathews correlation coefficient =", mcc ,"\n")
  cat("LaTEX:  &",specify_decimal(sens,2) ,"&",
         specify_decimal(spec,2), "&", specify_decimal(ppv,2) ,
         "&", specify_decimal(npv,2) ,"&", specify_decimal(corAll,2),
         "&", specify_decimal(mcc,4),"\\\\","\n")
  comp0frame <- as.data.frame(comp_0)
  comp1frame <- as.data.frame(comp_1)
  val_frame <- merge(comp0frame,comp1frame,all=TRUE)
  val_text <- data.table(val_frame,key="real_vals")
  val_text[.(0),text_val := "Non-fish-bearing"]
  val_text[.(1), text_val := "Fish-bearing"]
  dist_plot <- ggplot(val_text, aes(x=fit_vals,fill=text_val))
  dist_plot <- dist_plot + geom_density(alpha=.6)
  dist_plot <- dist_plot + labs(title="Predicted Probability
```

```
        Distributions for \nFish-Bearing and Non-Fish-Bearing Streams",
        x="Predicted Probability",y="Frequency",fill="Fish Presence")
  dist_plot <- dist_plot + theme(plot.title=element_text
        (family="cmr10",face="bold"))
  dist_plot <- dist_plot + theme(axis.text=element_text(family="cmr10"),
        axis.title=element_text(family="cmr10"))
  dist_plot <- dist_plot + theme(legend.position=c(.8,.8),
        legend.title=element_text(family="cmr10"),
        legend.text=element_text(family="cmr10"))
  print(dist_plot)
}
```

Appendix C

Input Data

C.1 Data set used for model development

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
10073	FALSE	0.231205518	4	12	55.01019298
10078	TRUE	0.088758273	1.666666667	12	55.03067071
10003	FALSE	0.039614992	6.666666667	12	55.17605
10016	FALSE	0.118240139	1	12	54.93201942
10017	FALSE	0.254497054	1.666666667	12	54.93103992
10027	FALSE	0.047441588	1	12	54.89540743
10029	FALSE	0.056281151	2.333333333	12	54.89390163
10049	FALSE	0.14431757	6.5	12	54.89005
10051	FALSE	0.62162624	1.75	12	54.86448566
10059	FALSE	0.102198474	7	12	54.81599685
10082	FALSE	0.441873504	2.2	12	55.16872522
10096.5	TRUE	0.02220776	7.666666667	12	55.33149478
10099	TRUE	0.176457627	6.333333333	12	55.33886267
10106	FALSE	0.924469615	1.5	3	55.31804498
10107	FALSE	0.223751354	4.2	12	55.31713754
10108	TRUE	0.103391377	9.666666667	12	55.32713634
11001	FALSE	0.486589417	2	12	55.35835449
11003	FALSE	0.833375887	12	3	55.35705364
11013	FALSE	1.946940466	0.036666667	3	55.50761738
11084	TRUE	0.027558932	9.25	3	54.19752934
1113	TRUE	0.884506367	1	3	55.05153149
1116	TRUE	20.29179772	2.166666667	3	55.04245343
1121	TRUE	1.710595198	5.9	3	55.03509757
12016	FALSE	0.996018294	0.416	3	55.61141432
12027	TRUE	1883.131427	0.021666667	3	55.51605914
12050	FALSE	0.335650244	2.4	3	55.90831995
12063	TRUE	4.959495319	2.166666667	3	55.36058685
12065	FALSE	1.770192095	12.66666667	3	55.33777904
12096	TRUE	5.719715615	5.833333333	3	55.93743507
12105	FALSE	0.04374398	8	12	55.5729685
12111	FALSE	0.025343094	8.833333333	12	55.63123524
12112	FALSE	0.170902846	6.5	12	55.63024628
12123	FALSE	0.289838632	14	3	55.50034926
12160	FALSE	0.064842562	12	12	55.20065
12161	TRUE	0.325961107	11.75	12	55.20026
12166	FALSE	0.643672745	2	3	55.2106
13010	TRUE	6.600715039	2.75	3	55.27741
13011	FALSE	0.262738858	8.5	12	55.27847
13012	FALSE	0.086737683	0.5	12	55.27001

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
13028	TRUE	6.532549019	1.333333333	3	55.04748361
13037	FALSE	1.653146734	1	3	55.00866357
13042	TRUE	3.323883862	1.066666667	3	55.13876799
13043	TRUE	262.1316658	1.833333333	3	55.29795282
13046	TRUE	17.73858212	7	3	55.3887145
13051	TRUE	0.760923867	1.75	3	55.41757145
13055	FALSE	4.595003331	25.83333333	3	55.45051244
13069	FALSE	0.060863898	1.6	3	55.56218465
13078	TRUE	1.719239977	10.5	3	55.5922695
13086	TRUE	37.19228144	5	3	55.2744304
13091	TRUE	0.224619928	8.833333333	3	55.21417005
13093	TRUE	570.6359972	9.166666667	3	55.210488
13102	TRUE	1.692494382	9.5	3	55.59693982
13104	FALSE	0.092007548	25	12	55.55718201
13110	FALSE	0.05934385	26	12	55.48852647
13121	TRUE	39.816289	2.833333333	3	55.6392056
13142	FALSE	1.269900081	2.666666667	3	54.89768675
13149	FALSE	0.105110563	1	12	54.89989133
13155	FALSE	1.169616427	3	3	54.83076646
1383	TRUE	0.108006683	12.6	3	54.25649
1384	FALSE	0.185570163	22	3	54.25388132
1385	FALSE	0.02630257	38.33333333	3	54.25210684
1386	FALSE	0.035380791	32.5	3	54.25140276
1388	TRUE	0.307363102	32.66666667	3	54.24950042
1399	TRUE	0.099212147	10	3	54.24149466
14003	FALSE	0.09179554	11.66666667	3	54.26227523
14005	TRUE	2.442045216	8.666666667	3	54.26551094
14007	TRUE	0.201342295	9.5	3	54.26840672
14032	TRUE	1.815287379	6.333333333	3	54.88863664
14036	FALSE	0.100002563	18	3	54.88845
14039	FALSE	0.197060376	40	3	54.884
1404	TRUE	0.015899757	20	12	54.23976
14043	FALSE	0.201238755	30	3	54.87447
14047	FALSE	0.471232171	4.333333333	3	54.86843648
14057	FALSE	0.167811712	12	3	54.876
14068	FALSE	0.471810559	20.75	3	54.8841
1411	FALSE	0.02153135	21.25	3	54.23576
1412	FALSE	0.026700021	25.6	3	54.23492
15005	TRUE	5.583185998	2	3	54.98062913
15007	TRUE	1.855202195	8	3	54.29202927
15008	TRUE	0.362197255	6	3	54.29178699

Site ID	Fish Bearing	Watershed	Gradient	Max. flow ≥ 80th	Latitude
15009	TRUE	2.061588776	15	3	54.2898663
15015	TRUE	0.173589673	20.5	3	54.28347171
15017	TRUE	0.037928833	17	3	54.27123
15018	FALSE	0.041995759	19.33333333	3	54.26159913
15023	TRUE	1.018114351	3.6	3	54.82209095
15025	TRUE	2.193666051	12.5	3	55.34109182
164.5	FALSE	0.006692367	7.166666667	12	55.56229497
165	FALSE	0.090649912	11.5	12	55.5628852
17	TRUE	1.724072272	23	3	55.9653392
17002	TRUE	10.32169173	7	3	55.63964321
17010	FALSE	0.378454702	0.042	3	55.67175577
17016	FALSE	0.393553715	0.06	3	55.6635504
17021	TRUE	7.693789956	0.89	3	55.63620858
17040	FALSE	1.539833938	0.19	3	55.61926844
17042	FALSE	3.505426907	1.381666667	3	55.59529158
17043	FALSE	0.463822006	0.064	3	55.60160276
17052	FALSE	1.576100857	0.125	3	55.56441083
17053	FALSE	2.033703659	0.03	3	55.55186981
17057	TRUE	0.117879823	0.02	3	55.50018446
17059	FALSE	0.255660049	0.055	3	55.48237318
17064	FALSE	1.793185908	14.4	3	55.04949
17066	TRUE	231.1685278	4.333333333	3	55.05804613
17069	TRUE	8.082517343	3	3	54.88763496
17073	FALSE	2.485606569	2	3	54.88735043
17079	FALSE	0.051742275	11	12	55.6968394
17081	FALSE	0.348086044	1	3	55.69516721
17096	FALSE	0.182653753	0.02	3	55.5064168
173	FALSE	0.176080186	12	12	55.57933732
174	FALSE	0.368249296	36	3	55.57509492
174.5	FALSE	0.058212189	19	12	55.57175941
176	FALSE	5.888426634	18.75	3	55.56674089
18	TRUE	3.305480988	4	3	55.96182108
181	TRUE	0.340182484	11	3	55.551728
185	FALSE	0.169779915	3.083333333	12	55.54586906
186	TRUE	2.094998415	3.083333333	3	55.54518443
189	FALSE	0.064920771	3.833333333	12	55.5439519
19004	FALSE	0.123615808	7	12	54.858565
19005	TRUE	0.429013147	3.833333333	12	54.84511872
19020	TRUE	0.819885642	4	3	55.33928513
19021	TRUE	0.203733383	4	12	55.33888407
19022	FALSE	0.779140792	2.25	3	55.23495

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
19023	FALSE	0.023684611	1.6	12	54.93193
19024	TRUE	1.600171542	1.5	3	55.34345595
19026	TRUE	0.230711501	3.5	12	55.33424413
193	FALSE	0.046458169	4.666666667	12	55.54310273
194	FALSE	0.094230055	5.333333333	12	55.54071894
196	FALSE	4.621013044	3.666666667	3	55.52819847
197	FALSE	0.21283923	3.666666667	12	55.52654855
199	FALSE	0.205448827	6.666666667	12	55.52453427
21001	FALSE	0.261539756	12.33333333	3	54.94896846
213	FALSE	0.655601985	11.33333333	3	55.48977321
215	FALSE	0.633346558	9	3	55.48346059
216	TRUE	0.629759837	2	3	55.47927933
222	TRUE	22.35447466	7	3	55.47416617
223	FALSE	0.265190252	10	3	55.47099799
224	FALSE	0.028357934	31	12	55.47006971
226	TRUE	0.368456951	15.16666667	3	55.46637492
229	FALSE	0.200990378	29.66666667	12	55.46330605
237	TRUE	3.356318681	10.16666667	3	55.43551744
239	TRUE	0.511513737	7.833333333	3	55.42961631
24	TRUE	2.207823696	3.75	3	55.92335599
254	FALSE	0.276176645	4	12	55.36169438
279.5	FALSE	0.065668171	14.33333333	12	55.36231579
285	FALSE	0.481370046	13.16666667	12	55.36374273
3	TRUE	45.44769693	1	3	56.15795082
3010	FALSE	0.873602433	4	3	55.32381224
3011	TRUE	7.289838488	3.833333333	3	55.3158405
3030	FALSE	0.066262089	6.8	12	55.22093
3039	TRUE	4.50067262	2.083333333	3	55.20404
3041	TRUE	5.422905477	4.5	3	55.18263
3044	TRUE	46.7431515	1	3	55.17551
3046	TRUE	0.923876343	3.5	3	55.14222216
3050	TRUE	85.99803936	0.833333333	3	55.12048396
3070	TRUE	0.77894088	2.2	3	54.99184364
3073	FALSE	3.240621822	1.666666667	3	54.98238599
3076	TRUE	5.766108154	1.333333333	3	54.95774418
3077	TRUE	31.67389708	1	3	54.95706497
3078	TRUE	54.81084924	1.2	3	54.952424
3080	TRUE	1.232057742	1.2	3	54.9496921
3081	FALSE	0.819423052	1	3	54.9469265
3089	FALSE	0.1551122	3.666666667	12	54.9230478
3112	TRUE	1.565972942	4.666666667	3	54.89662694

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
3115	FALSE	0.206357088	10	12	54.89126608
3117	FALSE	3.23200935	2.5	3	54.89600013
3123	TRUE	1.339069466	3.833333333	3	54.89847156
3124	TRUE	0.376163072	7.75	12	54.89845575
3151	TRUE	14.22625408	2.833333333	3	54.88794601
3170	TRUE	8.341749966	4.333333333	3	54.88462
3171	TRUE	8.195290051	3.333333333	3	54.88544
3172	TRUE	1.65204921	3	3	54.88569
3175	TRUE	6.494999428	2.5	3	54.88705
3176	FALSE	1.403399679	12.2	3	54.89090536
3179	FALSE	0.062306512	25	12	54.87827615
3180	FALSE	0.401552633	21.66666667	12	54.87557849
3181	TRUE	2.119171367	3.25	3	54.87264251
3192	FALSE	4.562558935	4	3	54.84299779
3196	TRUE	29.84193082	2.833333333	3	54.82878817
3198	TRUE	5.591327031	3.666666667	3	54.82323
3199	TRUE	5.599494949	2.666666667	3	54.82206
3203	FALSE	2.728745009	5	3	54.81961482
3205	TRUE	4.868032303	3	3	54.82273073
3206	FALSE	1.793651713	3.5	3	54.82451166
3211	TRUE	8.782020961	6.166666667	3	54.83084929
3213	TRUE	123.2370388	2.5	3	54.84591898
3217	TRUE	98.76892774	2.666666667	3	54.86730081
3226	FALSE	1.116370986	15.16666667	3	54.88971281
3229	TRUE	1.949264999	3	3	54.90827485
3231	TRUE	2.324703921	10	3	54.92510027
3235	FALSE	1.180705946	9	3	54.96225509
3237	TRUE	22.34637477	2	3	54.97293157
3238	FALSE	1.7687554	2.6	3	54.98234946
3239	FALSE	0.195202467	4.5	12	54.99175248
3240	TRUE	362.357308	1.333333333	3	55.02750448
3241	TRUE	9.493147191	1.8	3	55.03204023
3246	FALSE	1.744651823	2	3	55.07189
3248	TRUE	8.187585228	1	3	55.08450422
3264	TRUE	0.449994286	2.5	12	55.1524967
3267	TRUE	4.950499457	3.833333333	3	55.17235937
3269	TRUE	4.505121074	3	3	55.197177
3271	TRUE	4.408240615	2.833333333	3	55.22251157
3275	FALSE	2.906410703	3	3	55.23527304
3276	TRUE	0.877194674	11.66666667	3	55.24004748
3281	TRUE	2.19667065	5.75	3	55.2537892

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
3284	TRUE	1.029584496	6.2	3	55.26130612
3294	TRUE	4.983433893	1.333333333	3	55.28800225
3299	TRUE	27.43887993	2.833333333	3	55.30277546
3302	TRUE	4.875088329	3.8	3	55.31343303
3304	TRUE	0.733688191	3.75	3	55.3193055
3323	TRUE	0.193812502	6.166666667	12	55.33415456
3325	TRUE	0.651207961	15	3	55.33422001
3328	FALSE	0.972023782	4	3	55.33144661
3331	TRUE	6.111125105	8	3	55.33077871
3335	TRUE	0.115828857	14.5	12	55.33588482
3336	TRUE	0.619213044	8.333333333	12	55.33670756
3337	TRUE	0.27218175	5.833333333	12	55.33902862
3338	TRUE	0.680937419	11	3	55.34227029
3339	TRUE	15.89453116	8.166666667	3	55.34456871
3340	TRUE	1.357414147	14.33333333	3	55.3455255
3341	TRUE	0.499429305	15.83333333	12	55.34529686
3342	TRUE	1.30734534	6.833333333	3	55.34289626
3344	TRUE	12.94567658	8.5	3	55.34086125
3346	FALSE	7.952374671	11.33333333	3	55.33776923
3347	FALSE	0.333717528	5.166666667	12	55.33165244
3348	FALSE	2.449664548	10.16666667	3	55.33089227
3350	FALSE	0.964615766	9	3	55.32899839
3352	FALSE	1.943195088	10.6	3	55.32621518
3354	FALSE	0.432027115	15	12	55.32092422
3355	FALSE	2.528580691	1.833333333	3	55.31766081
3357	FALSE	0.469748891	1	12	55.32505331
3358	FALSE	0.94949245	13.83333333	3	55.32723395
3362	FALSE	1.183601163	36.4	3	55.32678805
3369	FALSE	0.315743771	25	12	55.32597
3370	FALSE	0.156904765	10	12	55.32687
3372	FALSE	0.100157316	21.66666667	12	55.32993
3373	TRUE	2.59377475	15.6	3	55.33295
3374	FALSE	2.017339217	11.66666667	3	55.33551
3376	FALSE	0.450726797	11.5	12	55.34372
3379	FALSE	0.920838401	6.5	3	55.34966
3379.4	FALSE	0.479193028	7	12	55.3517
3380	TRUE	4.042073137	9	3	55.35389
3381	TRUE	1.100520165	9.2	3	55.35677
3383	FALSE	0.500055267	15.5	3	55.3663
3384	FALSE	0.320796265	15.33333333	3	55.37077
3385	TRUE	0.444030829	6.333333333	3	55.3742

Site ID	Fish Bearing	Watershed	Gradient	Max. flow ≥ 80th	Latitude
3388	TRUE	197.6923856	3	3	55.38032
4	TRUE	1.152020366	1	3	56.15649758
5	TRUE	0.530909391	2	3	56.15047226
5016	TRUE	1.218017769	14.16666667	3	54.9712088
5017	TRUE	0.563834032	2.5	3	54.97278049
5018	TRUE	0.667722319	11.25	3	54.97006028
5019	TRUE	5.29866377	4.166666667	3	54.968332
5020	TRUE	4.437519438	6.5	3	54.96774301
5031	TRUE	2.524934697	19	3	54.9490931
7010	TRUE	6.929367779	2.2	3	55.87700648
7019	TRUE	2.794211509	3	3	55.81511129
7021	TRUE	3.08672999	3	3	55.79403684
7023	FALSE	0.538832035	4	3	55.79254885
7024	FALSE	0.081986191	7	12	55.79217537
7027	FALSE	0.534579159	5	3	55.78838949
7029	FALSE	0.42020534	5	3	55.77882
7030	TRUE	11.96857983	4.166666667	3	55.77900414
7037	TRUE	27.93942871	1.833333333	3	55.7657504
7040	TRUE	3.250724617	3.333333333	3	55.76678834
7041	TRUE	0.804990678	1	3	55.76637872
7062	FALSE	0.156862627	15	12	55.73659657
7067	FALSE	0.446042413	8.666666667	3	55.72706489
7081	TRUE	227.2901423	1.916666667	3	55.70629627
7089	TRUE	0.830353297	12	3	55.69338385
7090	FALSE	0.370486758	15	3	55.69346862
7093	TRUE	47.88686368	3.333333333	3	55.69075335
7095	FALSE	0.12290382	10	12	55.68967234
7098	FALSE	1.056398101	12	3	55.68966114
7099	TRUE	124.0066896	2.25	3	55.69161505
7116	TRUE	21.13604623	7.166666667	3	55.70077879
7118	TRUE	45.26061262	1.833333333	3	55.69661727
7122	FALSE	0.242728487	11.5	12	55.68971447
7125	FALSE	1.260071976	8.166666667	3	55.68754794
7127	FALSE	1.52960854	9	3	55.68703846
7128	FALSE	0.58482688	11.83333333	3	55.68630792
7129	FALSE	0.157465299	13	12	55.68487601
7130	FALSE	0.168619309	31	12	55.68135358
7132	TRUE	0.172107224	2	12	55.67655931
7135	FALSE	0.13315371	25.33333333	12	55.6674319
7137	TRUE	0.022534045	16.8	12	55.66628563
7138	FALSE	0.027473598	38	12	55.66533789

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
7139	FALSE	0.120720017	15	12	55.66068134
7158	FALSE	0.12353361	9.4	12	55.62472367
7159	FALSE	0.054337608	9	12	55.62261068
7162	FALSE	0.030945316	28.33333333	12	55.61882698
7163	TRUE	0.928097598	12.6	3	55.61517462
7163.5	FALSE	0.047666967	10	12	55.61440581
7164	TRUE	0.163631779	14.83333333	12	55.612427
7165	FALSE	0.030603698	4.666666667	12	55.61091421
7166	FALSE	0.09698288	6.166666667	12	55.60964724
7182	FALSE	0.662918209	7.25	3	55.54346663
7182.5	TRUE	0.194757557	4.75	12	55.5461751
7183	TRUE	0.60076778	5.833333333	3	55.54861289
7187	TRUE	1.929363443	5.083333333	3	55.55671118
7191	FALSE	0.218457269	8.25	12	55.55799048
7201	FALSE	0.10969662	2	12	55.36394083
7202	FALSE	0.134198167	2	12	55.36460545
7203	TRUE	5.026663982	1	3	55.36658198
7204	FALSE	0.096958905	3.166666667	12	55.36773671
7206	FALSE	0.055603699	1	12	55.36644865
7213	FALSE	0.058986555	8	12	55.35380848
7219	FALSE	0.025587648	2.333333333	12	55.2646
7228	TRUE	35.41847202	4.5	3	56.11953
7242	FALSE	0.738452301	7.833333333	3	55.21715
7243	FALSE	0.634442924	5.833333333	12	55.21368
7343	TRUE	7.343960693	8.333333333	3	55.63715379
7347	TRUE	0.296370979	1	3	55.63424842
7360	FALSE	16.83741498	0.095	3	55.65178871
7451	FALSE	0.162027741	4.333333333	3	55.41575402
7453	FALSE	0.507789147	4.666666667	3	55.40460414
7486	TRUE	26.7137007	2.833333333	3	55.13387953
7525	FALSE	0.864551197	2.2	3	54.88821
7528	TRUE	0.804466293	2.5	3	54.8896
7549	FALSE	0.417576925	3.333333333	12	55.3748677
7550	TRUE	6494.575239	0.5	3	55.3731065
7553	TRUE	4.302076404	3.333333333	3	55.34871
7559	TRUE	0.592304346	11	3	55.38559
7560	TRUE	0.430737487	11	3	55.38559
7864	FALSE	4.298918525	31	3	54.9675965
8	TRUE	43.21123165	1.916666667	3	56.10882874
929	TRUE	1.459638438	1.666666667	3	55.63897678
930	FALSE	0.796506792	1	3	55.63879394

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
971	TRUE	32.0179543	7.5	3	55.6353187

C.2 Data set used for model testing

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
10	TRUE	30.35720065	1.333333333	3	56.05583929
10031	FALSE	0.446759056	1.666666667	12	54.89636828
10033	FALSE	0.031702743	4	12	54.89838479
10056	TRUE	1.711839354	2.5	3	54.83778849
10058	TRUE	0.119214749	2.2	12	54.82871197
10070	FALSE	0.239829391	10	12	55.00256231
10072	FALSE	0.089115065	3.5	12	55.00853778
10080	FALSE	0.205112328	1.5	12	55.16195461
10098	TRUE	0.365509296	2.2	12	55.33095814
10103	TRUE	0.261191013	5.333333333	12	55.34091333
10104	FALSE	0.060613905	5.75	12	55.3299439
11002	FALSE	0.097564625	14.375	12	55.35879752
11005	TRUE	0.188525387	6.333333333	12	55.35829999
11007	TRUE	1199.613873	2	3	55.62534808
1102	FALSE	0.284265268	6	3	55.12232743
11083	FALSE	0.00498539	15.5	12	54.19886048
1114	TRUE	9.640001497	1.333333333	3	55.04480604
1115	TRUE	1.001378347	10	3	55.04312503
1117	TRUE	1.03139976	17.5	3	55.04093232
1118	FALSE	0.643896826	27.5	3	55.04036605
1119	TRUE	1.334428429	10.5	3	55.03915833
1120	FALSE	4.679001716	60	3	55.03706721
12	TRUE	66.08692535	1.166666667	3	56.02716536
12003	TRUE	38.51437421	0.03	3	55.616573
12013	FALSE	1.136418339	0.0725	3	55.61063773
12049	FALSE	2.409833322	1.166666667	3	55.9096499
12052	FALSE	0.671310349	3	3	55.90350943
12053	FALSE	1.699721623	2.5	3	55.90082973
12062	TRUE	10.28224424	3.333333333	3	55.35454386
12097	FALSE	0.251524406	20	12	55.76626868
12098	FALSE	0.737550171	11	3	55.76466137
12100	FALSE	0.149381457	40	12	55.76277477
12101	FALSE	0.404847509	11	3	55.76361831
12102	FALSE	0.242100572	10	12	55.57685995
12103	FALSE	0.061962506	15	12	55.57667663
12104	FALSE	0.128056937	4	12	55.57330323
12106	FALSE	0.108063393	10	12	55.57028387
12108	FALSE	0.242244197	14.8	12	55.59278673
12109	FALSE	0.106568794	11.66666667	12	55.63465253

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
12110	TRUE	0.53060797	14.8	3	55.63235873
12113	TRUE	0.184008497	4.333333333	12	55.62914429
12114	TRUE	93.1443085	1	3	55.57852429
12115	TRUE	93.1443085	2.166666667	3	55.57853739
12119	TRUE	42.57358572	2.75	3	55.50693639
12120	FALSE	0.055764415	13.5	12	55.50585633
12121	TRUE	1.686179423	9.4	3	55.50404901
12124	TRUE	1.214354112	9.375	3	55.49682116
12157	FALSE	0.028384824	3	12	55.1999
12158	TRUE	1.998924679	6.583333333	3	55.19992
12159	FALSE	0.397065006	4.25	12	55.20057
12173	TRUE	0.36920942	1	12	54.89907
12174	TRUE	0.36999646	1.5	12	54.89901
12175	TRUE	28.58857015	1	3	54.89901
13	FALSE	0.356221255	4.6	3	55.99387
13039	TRUE	5.731502863	2	3	55.12443982
13040	TRUE	0.860762885	4.333333333	3	55.13157
13041	TRUE	18.20401173	2.166666667	3	55.13641911
13052	FALSE	1.40632108	11.33333333	3	55.42555936
13053	TRUE	3.550964394	7.166666667	3	55.42714734
13054	TRUE	6.695140887	18.66666667	3	55.43709461
13056	TRUE	38.90150315	10.5	3	55.45661656
13057	TRUE	23405.15066	1.25	3	55.46428108
13068	TRUE	1.773839088	1.666666667	3	55.55713139
13081	FALSE	0.812994169	2.25	3	55.62646325
13090	TRUE	0.326257839	4.166666667	3	55.21582562
13098	FALSE	0.084514652	20	12	55.77125683
13099	TRUE	0.239316811	7.25	12	55.64731213
13101	FALSE	0.11207399	32.66666667	12	55.60200816
13103	TRUE	1.409266021	6.2	3	55.59005375
13106	FALSE	0.537715388	13.66666667	3	55.557
13107	FALSE	0.061555395	15.08333333	12	55.51012288
13108	FALSE	0.080105469	26.5	12	55.50943282
13109	TRUE	78.17818381	2	3	55.49035868
13111	FALSE	0.024634005	20.4	12	55.47659849
13122	FALSE	0.069256472	5	3	55.62412244
13123	FALSE	0.706441399	70	3	55.03698149
13139	FALSE	1.219925986	5	3	54.89901289
13156	FALSE	0.050216301	0.5	12	55.21289067
13178	FALSE	0.119565083	9	12	55.36161033
1382	FALSE	0.243855991	15.66666667	3	54.25803

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
1396	TRUE	113.5880961	3.5	3	54.24817
1397	TRUE	0.042189474	7	3	54.24425
1398	TRUE	0.060994391	7.5	3	54.24203784
14	TRUE	72856.58032	1.5	3	55.98799871
14001	TRUE	4.301178595	6.8	3	54.26242528
14002	FALSE	0.072301655	11.8	3	54.26211705
14004	TRUE	1.779658305	2	3	54.26266757
14006	TRUE	0.056131051	7	3	54.26698768
14008	FALSE	0.128237882	17.4	3	54.268432
14011	TRUE	0.218087289	12.75	3	54.271289
14012	TRUE	0.044229624	30	3	54.27221
1403	TRUE	0.062006853	31.2	3	54.239862
14031	TRUE	4.992448182	2.75	3	54.88530317
14034	TRUE	70.65429071	2	3	54.89235723
14035	TRUE	0.048022252	11.5	3	54.8936
14038	FALSE	0.749084105	35	3	54.886
14040	FALSE	0.231482268	40	3	54.882
14041	FALSE	1.476807475	28.33333333	3	54.878
14044	FALSE	0.197424762	13.25	3	54.87373
14045	FALSE	0.132116404	34.5	3	54.54296
14046	FALSE	0.542636762	19.16666667	3	54.86698412
1405	TRUE	0.016268135	24.16666667	12	54.23794
14077	FALSE	0.841203551	4	3	55.38345495
15011	FALSE	0.693851383	17	3	55.70200775
16	FALSE	1.266554614	27.33333333	3	55.9689318
166	TRUE	127.0536284	1.833333333	3	55.56748422
168	FALSE	0.512139675	21.16666667	3	55.57187078
169	FALSE	2.670550716	13	3	55.57559337
17005	TRUE	5.930403504	6.333333333	3	55.64870217
17011	FALSE	0.258741266	0.03	3	55.67115661
17012	FALSE	4.18594137	0.048333333	3	55.67101966
17017	TRUE	14.38410183	0.035	3	55.65851433
17018	FALSE	0.504040143	0.041666667	3	55.65510667
17037	FALSE	0.322492053	0.02	3	55.62663398
17044	FALSE	1.419429552	0.1175	3	55.60233907
17049	FALSE	0.273213311	0.063333333	3	55.58307002
17050	FALSE	0.964088737	0.03	3	55.57837395
17061	TRUE	0.932901105	3.416666667	3	55.07475439
17062	TRUE	0.240498012	5.25	3	55.07995187
17063	TRUE	6.533886315	2.4	3	55.07924619
17068	FALSE	0.647451005	14.25	3	55.05737

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
17074	TRUE	4.481789175	2	3	54.88735361
17076	TRUE	9.677780091	3.5	3	55.57483895
17078	FALSE	0.053744904	2	12	55.69675599
17080	FALSE	0.046503579	3	12	55.69626331
17082	TRUE	0.041020893	10	12	55.35734564
17086	TRUE	5.908477738	1	3	55.10235111
17097	TRUE	0.873912601	1.5	3	55.05167055
174.4	FALSE	0.270282216	28.33333333	3	55.57214551
175	FALSE	0.815689447	28.5	3	55.57081383
180	FALSE	0.286916369	15.66666667	3	55.55786095
18013	TRUE	0.140325696	1	12	55.13204
183	FALSE	0.042728445	41	12	55.54733991
184	FALSE	0.391676071	2.875	3	55.54649276
187	TRUE	4.161995527	2	3	55.5440585
188	TRUE	0.125098717	4.5	12	55.54398192
19007	TRUE	0.660760488	6.5	3	54.85324416
19010	TRUE	2.85756787	2.25	3	54.85097434
19011	TRUE	27.28450652	3.166666667	3	54.84362218
19012	TRUE	760.8721744	2.166666667	3	55.30193743
19013	TRUE	760.8721744	2	3	55.3031365
19016	FALSE	0.138023575	15.6	12	55.63250494
195	FALSE	0.027242114	3.4	12	55.54027867
198	FALSE	0.647958753	6	3	55.52583619
200	FALSE	0.133441067	80	12	55.51659864
20003	TRUE	4.145659812	1.666666667	3	55.63797233
20005	TRUE	1.376739104	1	3	55.03308264
21002	FALSE	3.078454754	24.83333333	3	54.942063
227	FALSE	0.127542109	29.33333333	12	55.46540337
23	FALSE	1.138595432	3.833333333	3	55.9246248
230	FALSE	0.284473188	11.33333333	3	55.4616697
231	TRUE	0.163317079	28.66666667	12	55.46104276
232	FALSE	0.472405969	28	3	55.45728738
233	FALSE	0.231098966	36.4	12	55.45714492
235	TRUE	2.66269888	14.6	3	55.44754005
235.5	TRUE	0.298795594	18.5	3	55.44246411
236	TRUE	0.457176432	5.25	3	55.43633558
256	FALSE	0.102502302	3.666666667	12	55.35909704
272	FALSE	0.077334309	2.75	12	55.35654128
277	FALSE	0.174867531	6	12	55.3608989
280	FALSE	0.081052822	17.5	12	55.36242716
281	FALSE	0.050037862	17.5	12	55.362729

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
282	FALSE	0.111836797	15.83333333	12	55.36325653
284	FALSE	0.066185356	8.2	12	55.36433909
3043	TRUE	0.292298043	5	12	55.17751
3047	TRUE	0.689945289	5	3	55.13805199
3048	TRUE	0.407804021	4.666666667	12	55.13244337
3049	TRUE	96.97132034	1	3	55.13231078
3066	FALSE	0.042584872	0.585	12	55.00234514
3071	FALSE	0.423554255	2.333333333	12	54.98982965
3072	FALSE	0.153370955	3.333333333	12	54.98742436
3082	TRUE	1.293276282	2.833333333	3	54.93682312
3083	TRUE	2.468802647	3.5	3	54.93432088
3091	TRUE	4.863791088	1	3	54.90478248
3095	TRUE	6.750409119	1	3	54.8951965
3099	FALSE	0.634072534	10	12	54.8943
3100	TRUE	5.545299092	3	3	54.89843
3116	FALSE	0.3225745	2	12	54.89162209
3121	FALSE	0.121646402	8.4	12	54.89412123
3125	TRUE	0.63126637	5.25	12	54.89870856
3127	FALSE	0.080171943	1	12	54.89975329
3130	TRUE	59.89446315	1	3	54.89755101
3158	TRUE	1.676026296	2.5	3	54.88898332
3163	TRUE	7095.912599	1	3	54.88298
3167	TRUE	9.891339818	5	3	54.88156
3168	TRUE	0.31659456	1.666666667	12	54.8828
3169	TRUE	0.789390849	6.833333333	3	54.88409
3173	TRUE	6.53831488	3	3	54.8858
3174	TRUE	6.521671066	3	3	54.88648
3178	FALSE	1.633441213	10.2	3	54.88019242
3193	TRUE	4.425382085	3.25	3	54.8416445
3197	TRUE	5.20228108	3.666666667	3	54.82634657
3201	FALSE	0.134751996	3	12	54.81573723
3207	TRUE	0.67533362	5.333333333	3	54.82725002
3209	TRUE	0.789023663	5.5	3	54.8294007
3214	FALSE	0.21870429	9	12	54.8477151
3220	TRUE	1.472198323	3	3	54.87573433
3223	TRUE	7.405647345	3	3	54.88165117
3227	TRUE	6.435058167	2	3	54.89462013
3232	TRUE	0.997695708	1	3	54.93618021
3233	TRUE	10.89200034	4	3	54.94114473
3234	FALSE	0.286209762	1	12	54.94551114
3244	TRUE	47.39526706	25.75	3	55.0652134

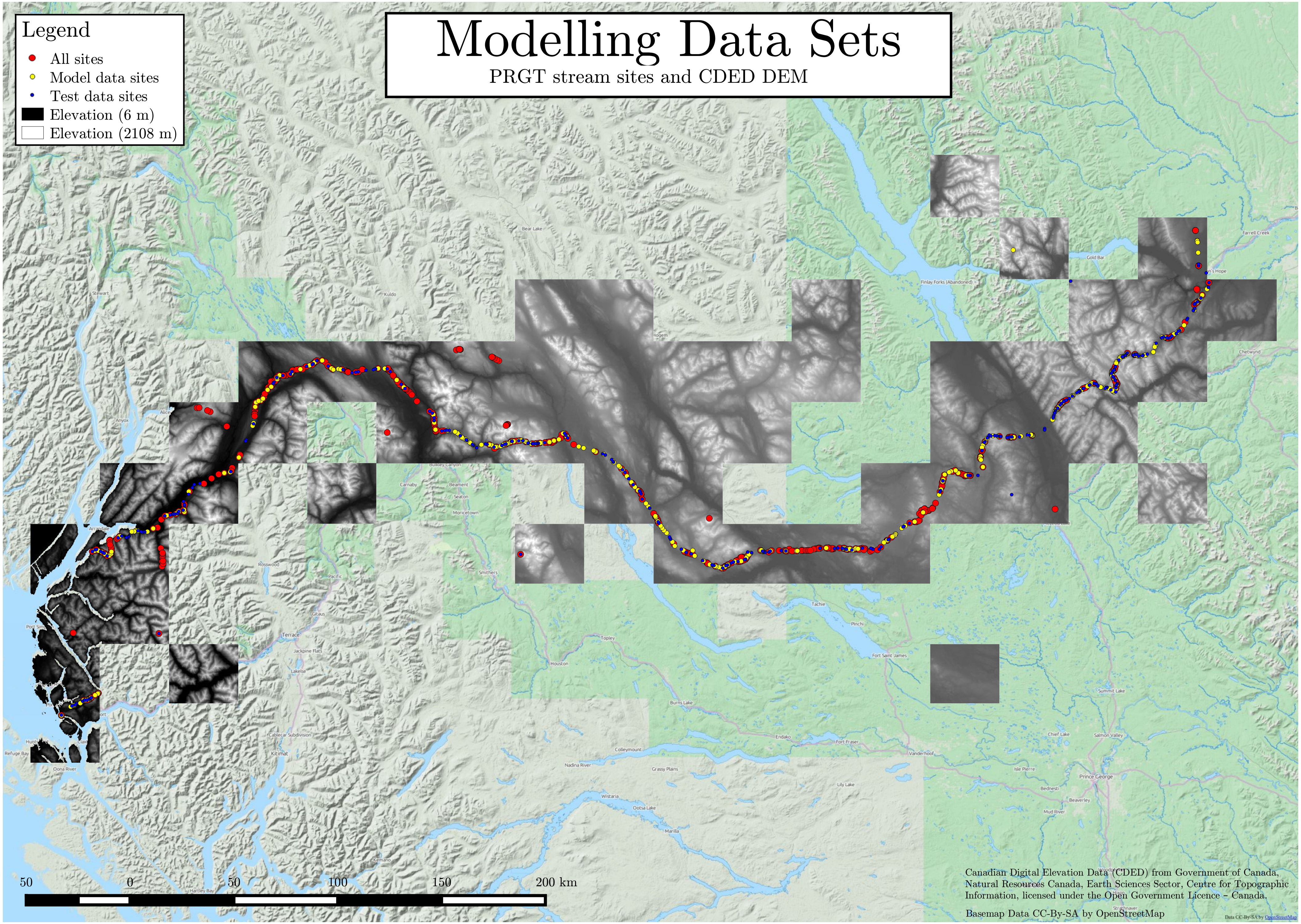
Site ID	Fish Bearing	Watershed	Gradient	Max. flow ≥ 80th	Latitude
3245	TRUE	47.39526706	1	3	55.0655738
3268	FALSE	0.105022378	7.5	12	55.18705503
3273	TRUE	0.86178188	3.333333333	3	55.22320265
3282	FALSE	0.723310936	6	3	55.25480316
3283	TRUE	1.030763254	6.8	3	55.25924913
3285	FALSE	0.41007205	3	12	55.26575095
3293	TRUE	1.946537479	3	3	55.28466188
3295	TRUE	7.555194549	3.4	3	55.28871537
3322	FALSE	0.180292271	9.666666667	12	55.3353835
3326	TRUE	5.240393527	6.333333333	3	55.33251979
3326.5	TRUE	0.015632689	4.333333333	12	55.33166549
3327	TRUE	0.789888415	4.4	3	55.33155735
3327.5	TRUE	0.00343601	4	12	55.33153504
3329	TRUE	1.850832914	7.5	3	55.3313391
3330.5	TRUE	0.027752198	5.75	12	55.3310645
3332	FALSE	0.225199082	16	12	55.33305944
3333	FALSE	0.059138215	10	12	55.33337767
3334	TRUE	1.796804501	10.66666667	3	55.33408039
3338.5	TRUE	0.023843181	10.6	12	55.34355857
3343	FALSE	0.316539618	12.16666667	12	55.34287789
3345	TRUE	2.514283797	11.66666667	3	55.34088475
3351	FALSE	0.50772298	1	12	55.32803102
3353	FALSE	2.821486799	15.33333333	3	55.32435341
3356	FALSE	0.586460702	3.8	12	55.31701107
3360	FALSE	1.947967017	17.75	3	55.32858086
3361	FALSE	0.69063888	17.83333333	3	55.32992628
3368	FALSE	0.315743771	70	12	55.32494
3371	FALSE	0.399826583	40	12	55.3279
3376.5	FALSE	0.080787153	7	12	55.34388
3377	FALSE	1.694086554	11	3	55.34519
3378	FALSE	0.360452472	9.5	12	55.34706
3382	FALSE	0.099248766	12.5	3	55.36368
3390	FALSE	0.339495937	21	3	55.38562
3392	FALSE	0.672753653	16.33333333	3	55.38511774
3553.5	FALSE	0.065069502	8.333333333	12	55.32378599
5015	FALSE	0.326669752	1.5	3	54.97145848
5016.25	FALSE	0.051189875	12	3	54.97129586
5021	TRUE	0.220358436	6	3	54.94952
5022	TRUE	0.222299443	6	3	54.94952
5023	FALSE	0.339383947	23	3	54.94925794
5024	FALSE	0.199925552	31	3	54.94767668

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
5025	FALSE	0.138801819	39	3	54.94511043
5026	TRUE	77.82694161	3	3	54.94130796
5027	TRUE	0.194234993	7.166666667	3	54.94250936
5028	TRUE	0.706908957	8.4	3	54.94334299
5030	TRUE	0.464032903	26.5	3	54.94966934
7006	TRUE	0.558701895	10.66666667	3	55.88351396
7006.5	TRUE	0.015684915	9.5	12	55.88320234
7011	TRUE	0.637193811	3.666666667	3	55.87574125
7011.5	FALSE	0.217857483	9	12	55.87464811
7022	FALSE	0.538832035	2	3	55.79332916
7025	FALSE	0.124069608	4	12	55.7903558
7028	TRUE	3.030616198	5.166666667	3	55.78108648
7039	TRUE	0.932294402	2.166666667	3	55.76683316
7042	TRUE	41.18519791	2.166666667	3	55.76453194
7056	FALSE	0.733408929	13.66666667	3	55.74195248
7087	FALSE	2.620722745	13.4	3	55.69751357
7091	FALSE	1.521839837	6	3	55.69297597
7092	FALSE	0.199878257	4	12	55.69173659
7096	FALSE	0.431704834	3	3	55.68975475
7097	FALSE	0.032424134	24	12	55.68975105
7101	FALSE	1.074343242	2	3	55.69583622
7102	TRUE	15.4428087	5	3	55.69602282
7107	FALSE	0.088666428	2.8	12	55.69393155
7108	FALSE	3.718512854	3.5	3	55.69354934
7117	TRUE	1.710279597	9	3	55.6978339
7119	FALSE	0.052529274	15	12	55.69379676
7120	FALSE	0.341229056	9	3	55.69339969
7121	FALSE	0.900986117	21.66666667	3	55.69187942
7123	FALSE	0.126901956	20	12	55.68807347
7124	FALSE	0.208106668	17	12	55.68790181
7126	FALSE	0.146970068	8.25	12	55.68716949
7133	TRUE	7.931380269	6.6	3	55.67398156
7143	TRUE	9.008307212	3.833333333	3	55.64626636
7161	FALSE	0.488155521	13.66666667	3	55.61988772
7167	FALSE	0.921101693	7.8	3	55.60680072
7177	TRUE	3.613749068	11	3	55.56142055
7178	TRUE	0.156227116	9.7	12	55.55969684
7179	TRUE	0.320171581	7.333333333	3	55.54344348
7186	TRUE	0.707672717	6.2	3	55.5516129
7188	FALSE	6.91217982	11	3	55.5620088
7189	FALSE	0.819266545	18.66666667	3	55.56165612

Site ID	Fish Bearing	Watershed	Gradient	Max. flow $\geq$ 80th	Latitude
7190	FALSE	0.106472895	11.6	12	55.55913989
7192	FALSE	0.368655428	10.66666667	3	55.55731226
7194	TRUE	82.07213164	2	3	55.39331
7195	FALSE	0.257553128	1	12	55.36662
7195.5	FALSE	0.026780347	1	12	55.38728
7204.5	FALSE	0.028125429	3	12	55.36600662
7207	FALSE	0.422147358	5.333333333	12	55.36699393
7208	TRUE	2.599072769	5.833333333	3	55.36296027
7211	FALSE	0.079223852	5	12	55.35746857
7212	TRUE	1.061092202	9.833333333	3	55.35446613
7217	TRUE	7.2911842	3.166666667	3	55.33926597
7218	FALSE	0.334534921	18	12	55.33649062
7226	TRUE	59.21724844	2.833333333	3	55.20044
7229	FALSE	1.628382399	11	3	55.12101
7331	TRUE	6.594801245	0.666666667	3	55.63705522
7332	FALSE	0.036449575	0.5	12	55.63709652
7335	TRUE	39.78524297	1.5	3	55.63175461
7366	FALSE	0.077453406	0.15	3	55.64794503
7452	FALSE	0.193896397	3.8	3	55.41294132
7455	TRUE	16.18259063	4.666666667	3	55.34675513
7456	TRUE	117.511165	3.333333333	3	55.34163044
7471	TRUE	19264.37587	2.833333333	3	55.16724851
7479	TRUE	19305.63496	2	3	55.16295018
7480	FALSE	2.284605772	0.5	3	55.15230309
7482	TRUE	12.46758036	1.25	3	55.14932338
7488	TRUE	62.66195744	3.833333333	3	55.09297454
7526	FALSE	0.896883999	2.333333333	3	54.88698
7546	TRUE	16.01991361	3.333333333	3	55.36530288
7548	FALSE	0.325604998	3.333333333	12	55.37369148
7556	TRUE	222.1437047	3.833333333	3	55.31443
7561	TRUE	0.387857174	11	3	55.38622864
7866	TRUE	4.04538659	25	3	54.9661538
7867	TRUE	1.839023099	7	3	54.96630399
7868	TRUE	0.630696644	9.5	3	54.96543644
7869	TRUE	36.04617057	7	3	54.96759649
9	TRUE	30.35720065	1.583333333	3	56.06266306
9016	TRUE	2.602809701	1	3	55.84278539
9017	TRUE	5.134000098	2.5	3	55.83750677
945	FALSE	66.44174413	0.5	3	55.6342519
959	TRUE	7.248440113	1	3	55.6218945
968	TRUE	4.938526421	1.166666667	3	55.63441282

Appendix D

Map of Data Sets


Modelling Data Sets
PRGT stream sites and CDED DEM
Legend
All sites
Model data sites
Test data sites
Elevation (6 m)
Elevation (2108 m)
50
0
50
100
150
200 km
Canadian Digital Elevation Data (CDED) from Government of Canada, Natural Resources Canada, Earth Sciences Sector, Centre for Topographic Information, licensed under the Open Government Licence - Canada.
Basemap Data CC-By-SA by OpenStreetMap

Appendix E

Predicted Probability Frequency Distributions

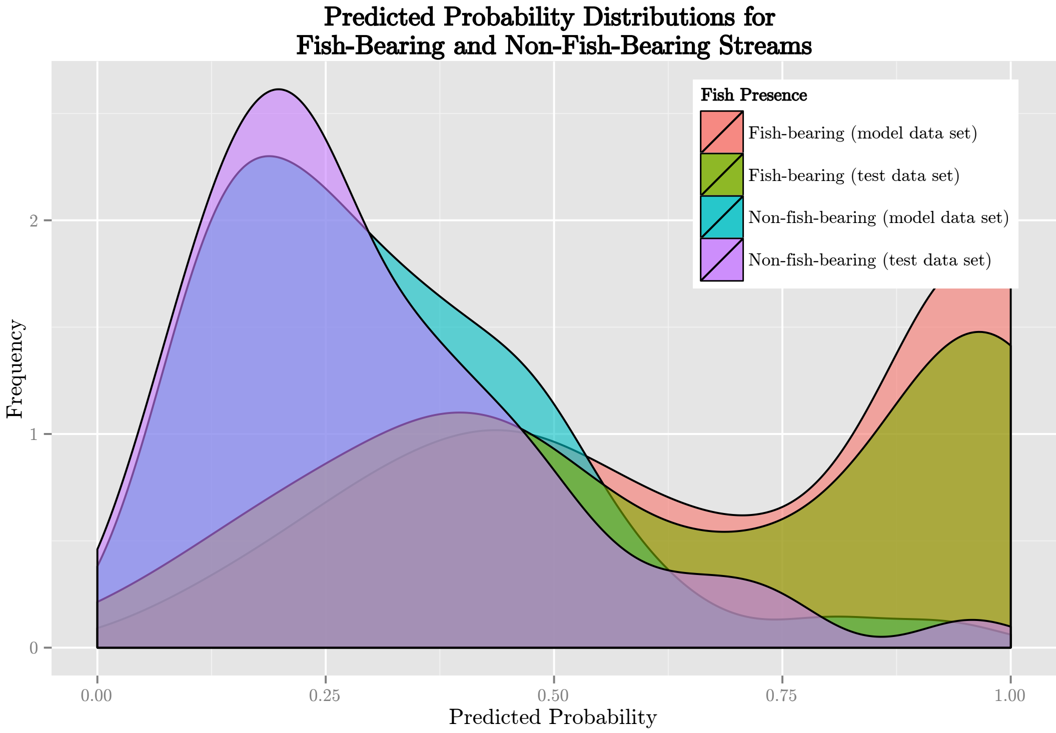


Figure E.1: Overlapping distributions of probability frequencies from model 1 (modelling and testing data sets).

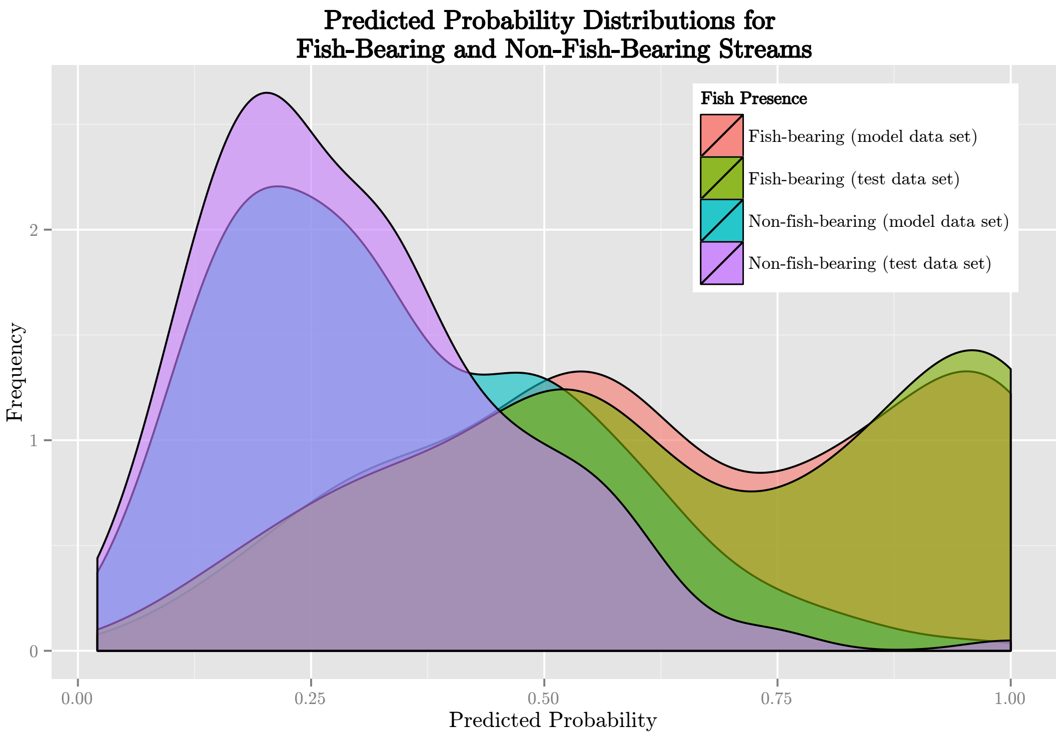


Figure E.2: Overlapping distributions of probability frequencies from model 2 (modelling and testing data sets).

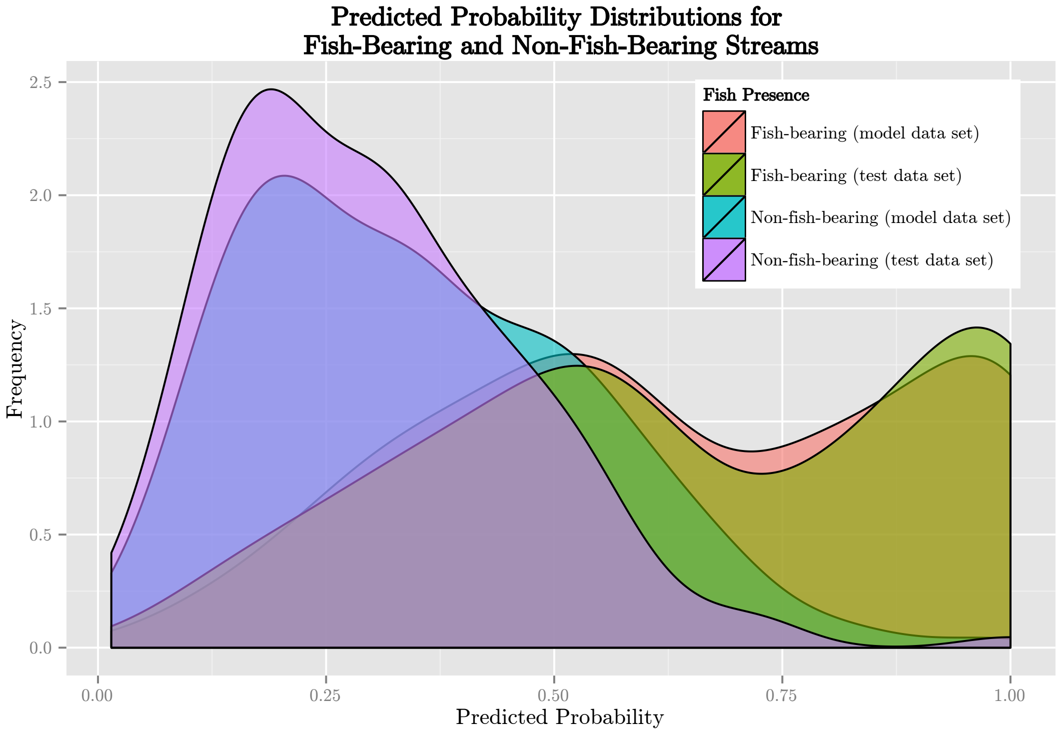


Figure E.3: Overlapping distributions of probability frequencies from model 3a (modelling and testing data sets).

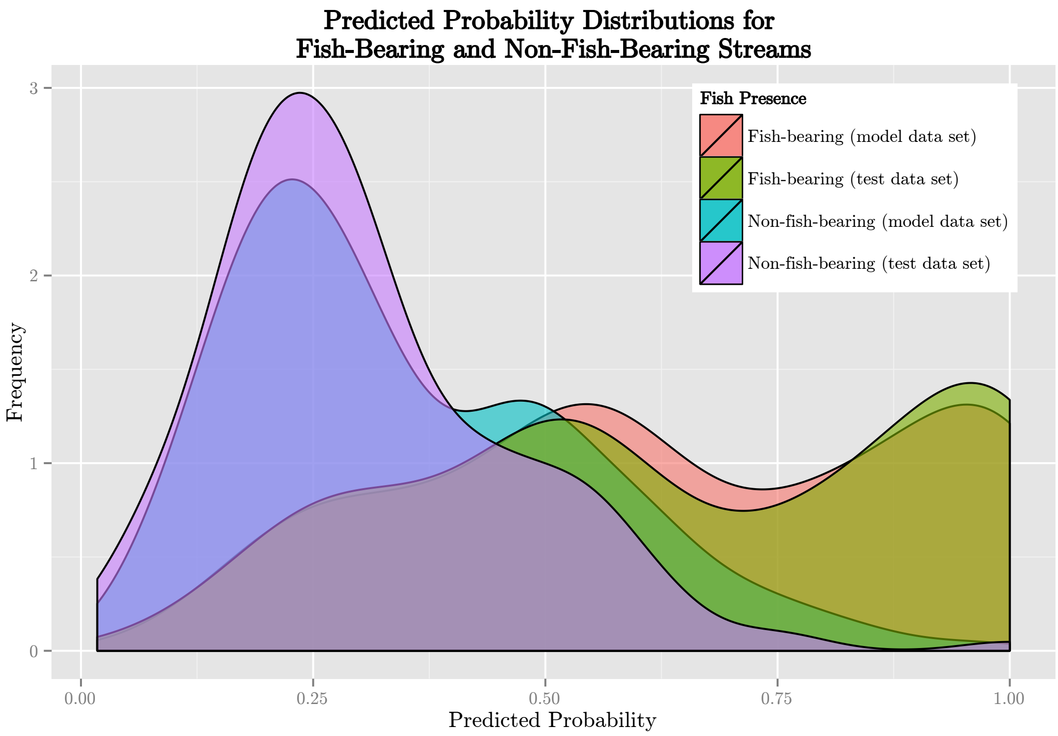


Figure E.4: Overlapping distributions of probability frequencies from model 3b (modelling and testing data sets).

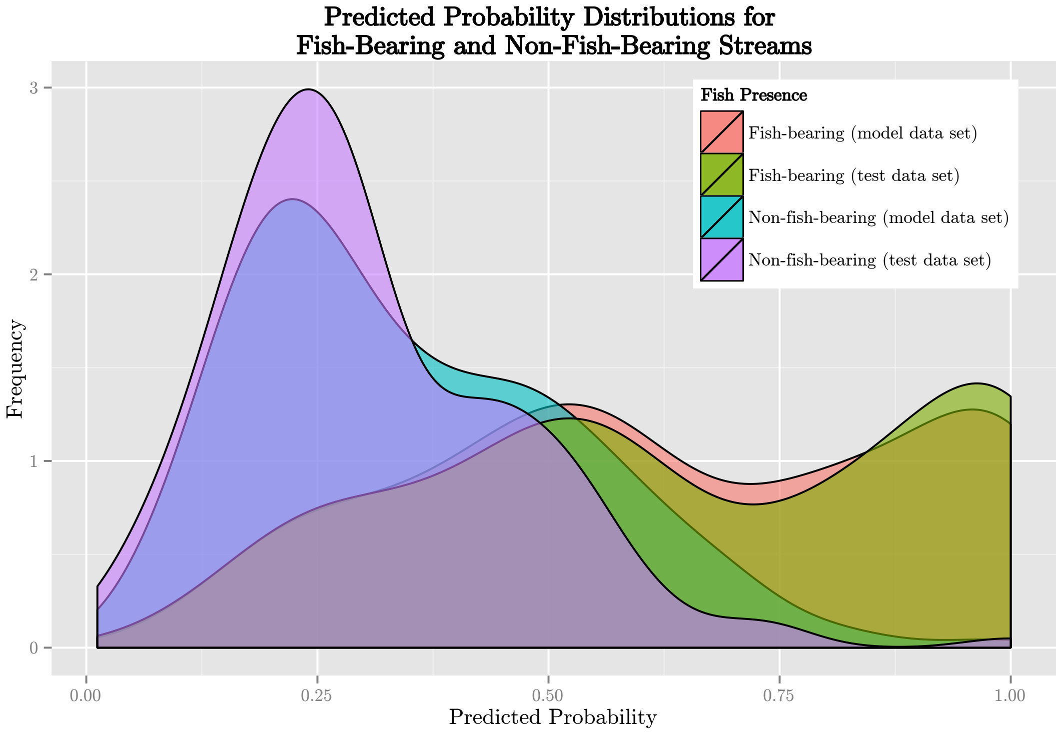


Figure E.5: Overlapping distributions of probability frequencies from model 4a1 (modelling and testing data sets).

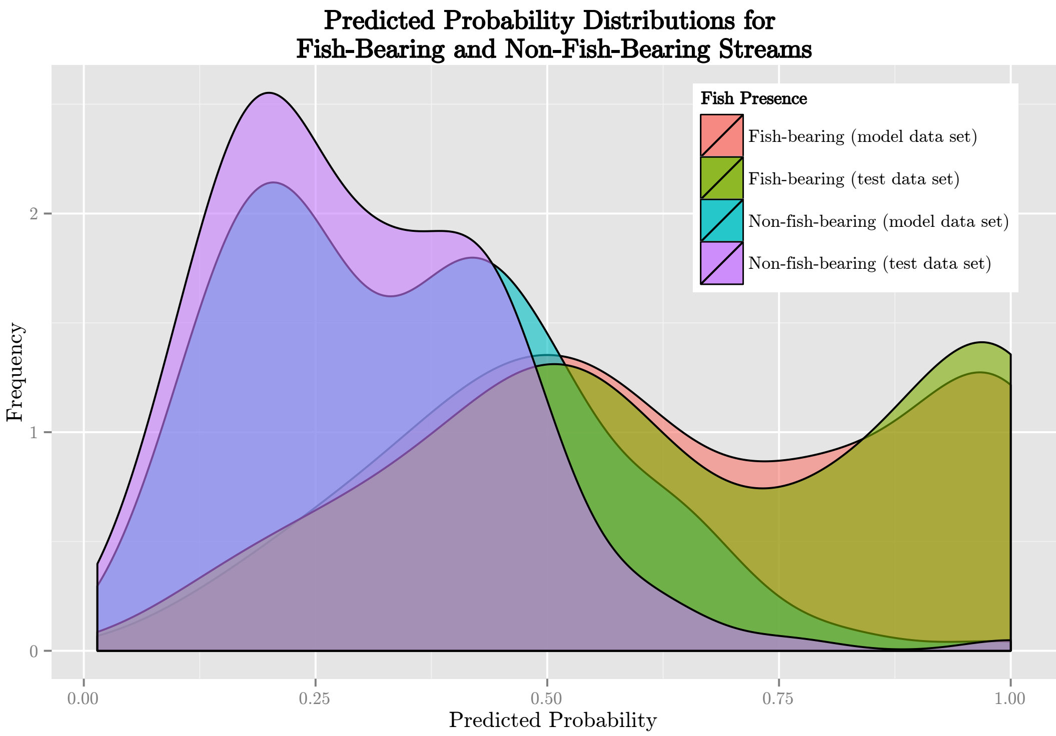


Figure E.6: Overlapping distributions of probability frequencies from model 4a2 (modelling and testing data sets).

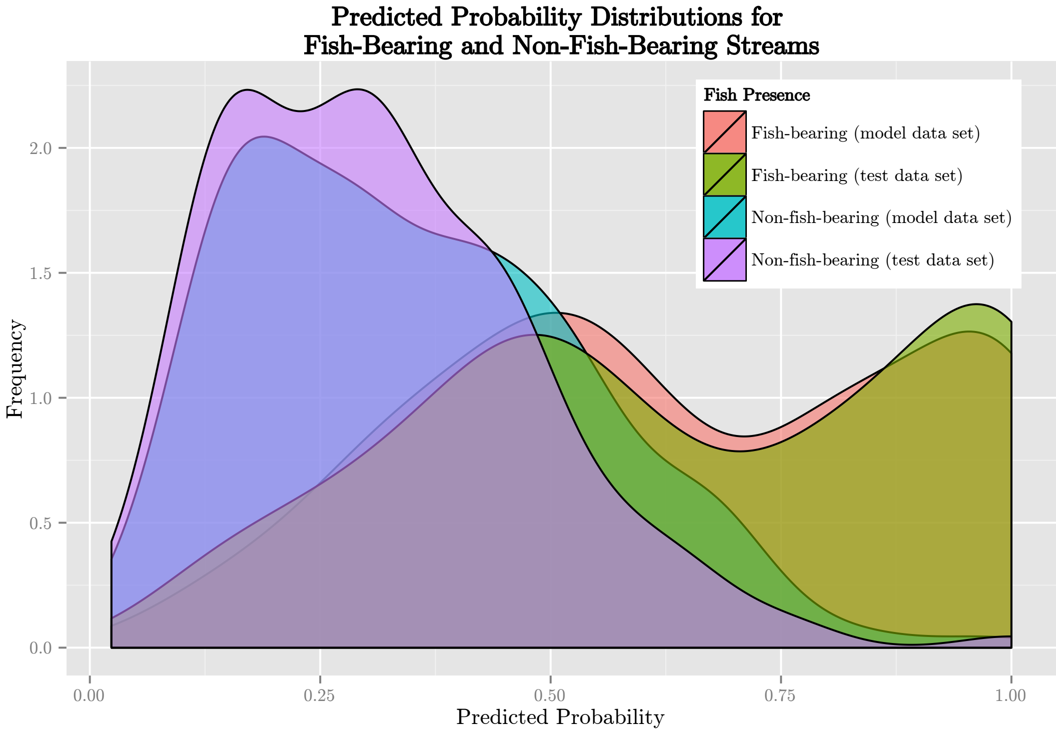


Figure E.7: Overlapping distributions of probability frequencies from model 4a3 (modelling and testing data sets).

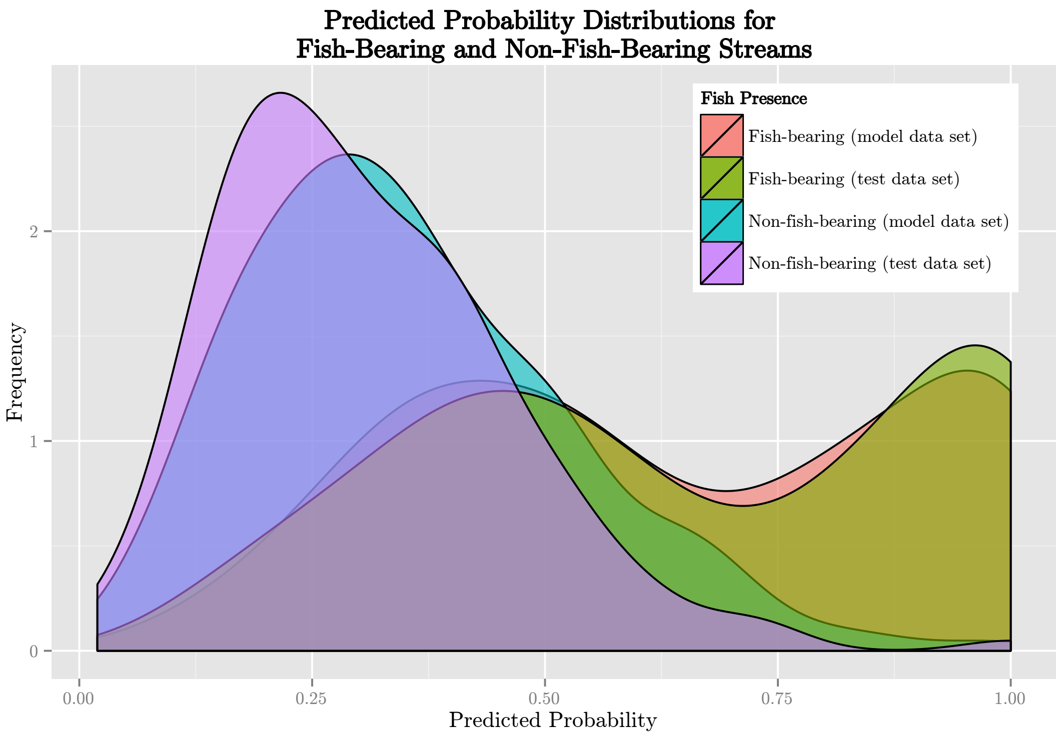


Figure E.8: Overlapping distributions of probability frequencies from model 4a4 (modelling and testing data sets).

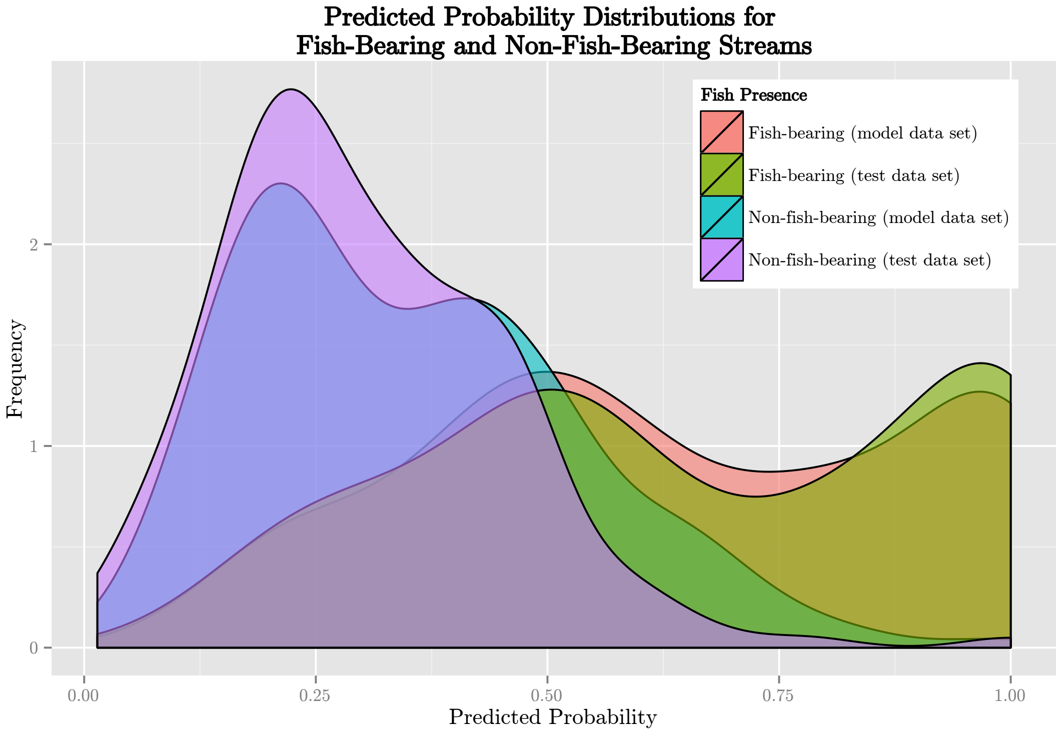


Figure E.9: Overlapping distributions of probability frequencies from model 4b2 (modelling and testing data sets).

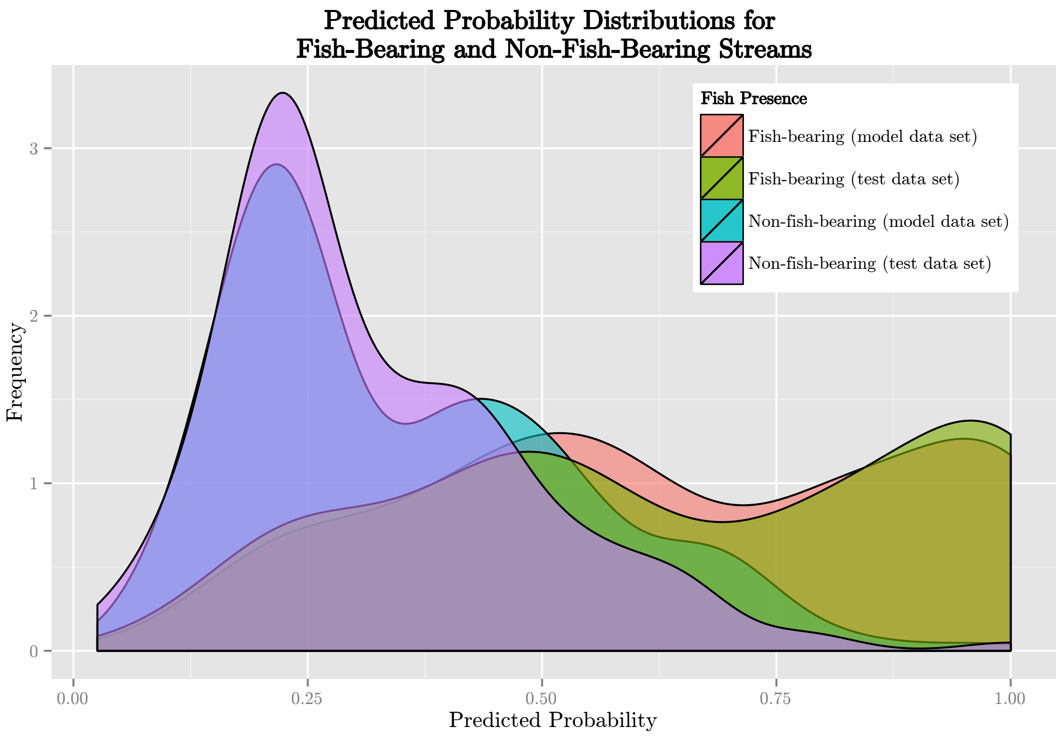


Figure E.10: Overlapping distributions of probability frequencies from model 4b3 (modelling and testing data sets).

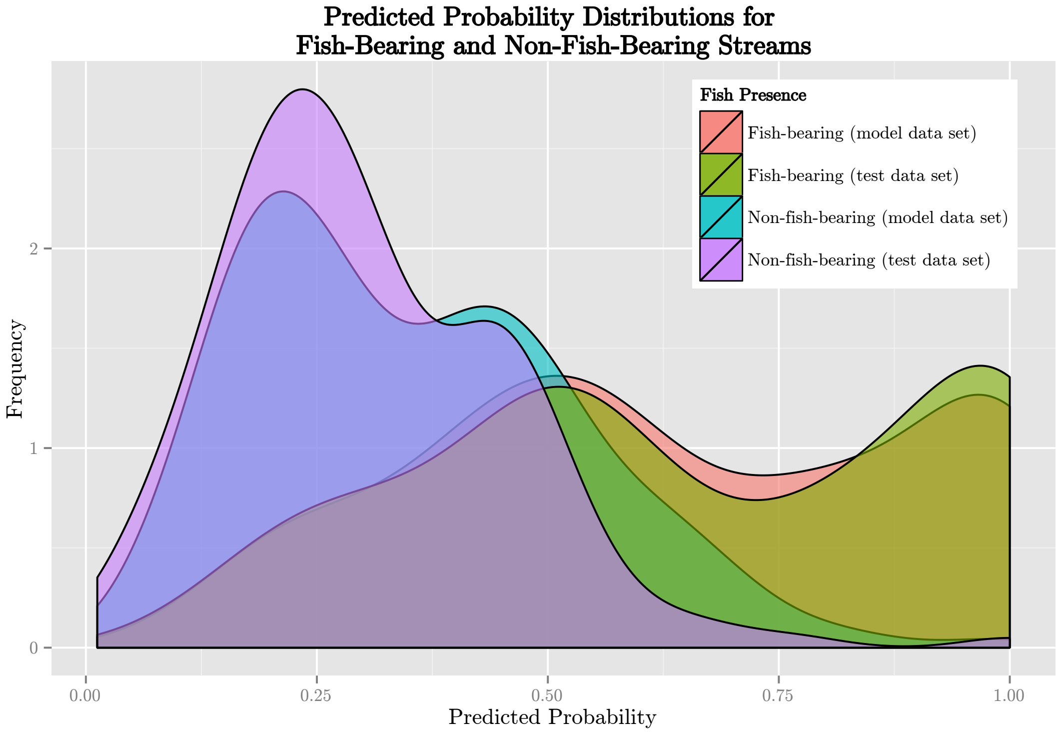


Figure E.11: Overlapping distributions of probability frequencies from model 5 (modelling and testing data sets).

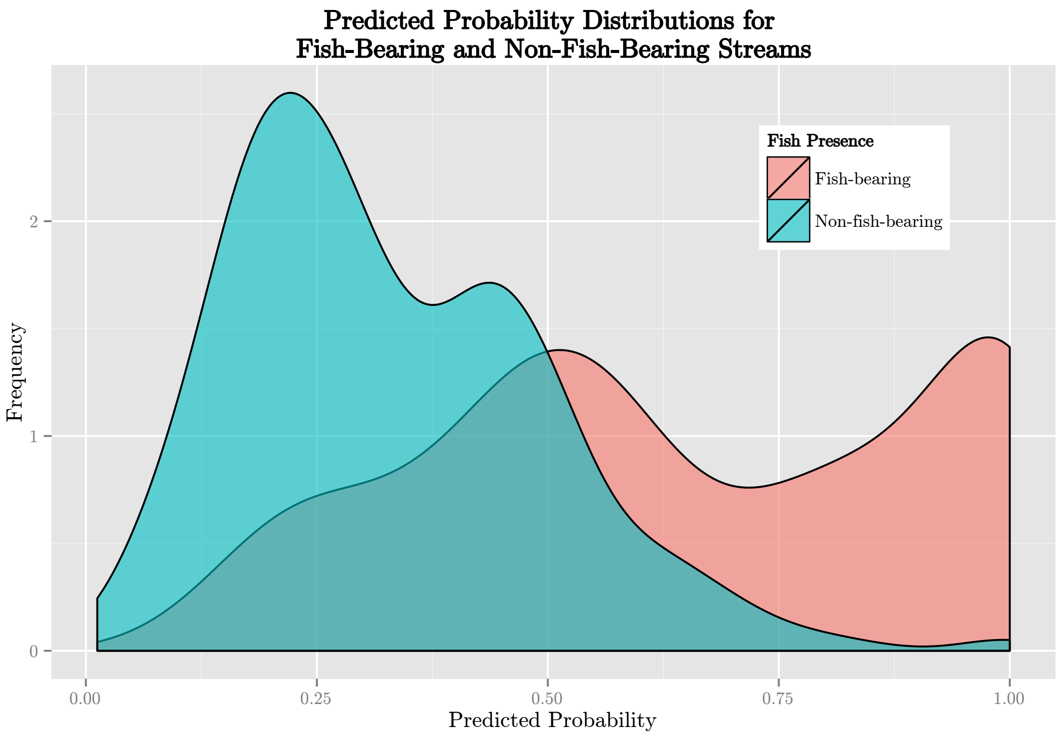


Figure E.12: Overlapping distributions of probability frequencies from model 5 (combined data set).

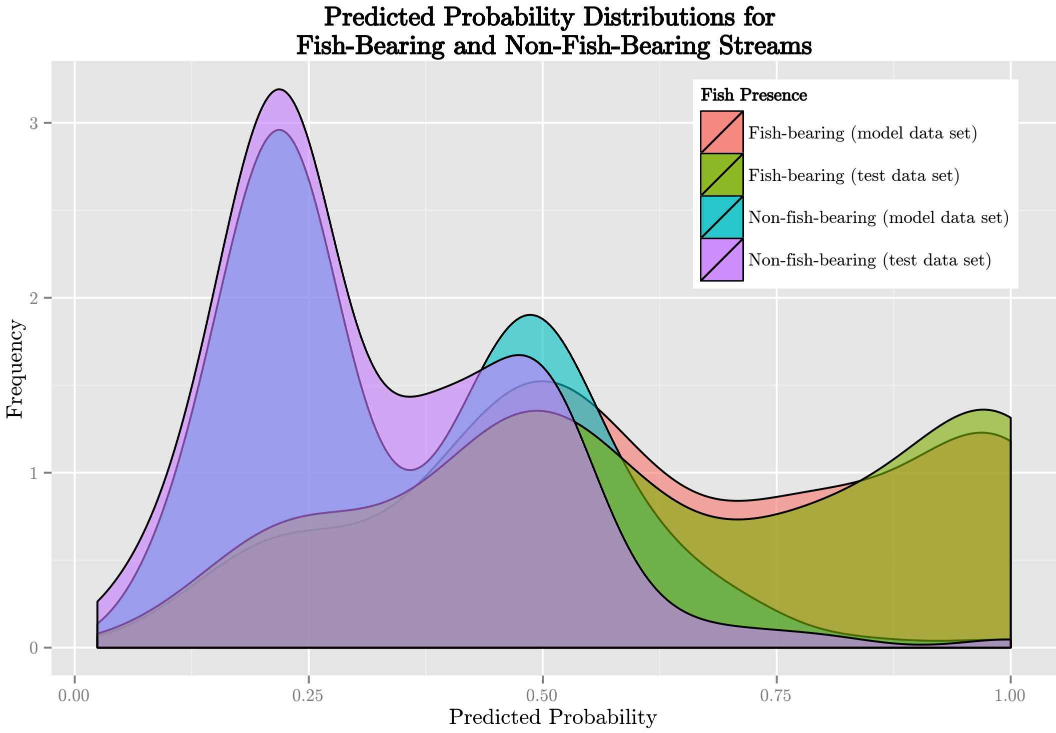


Figure E.13: Overlapping distributions of probability frequencies from model 6 (modelling and testing data sets).

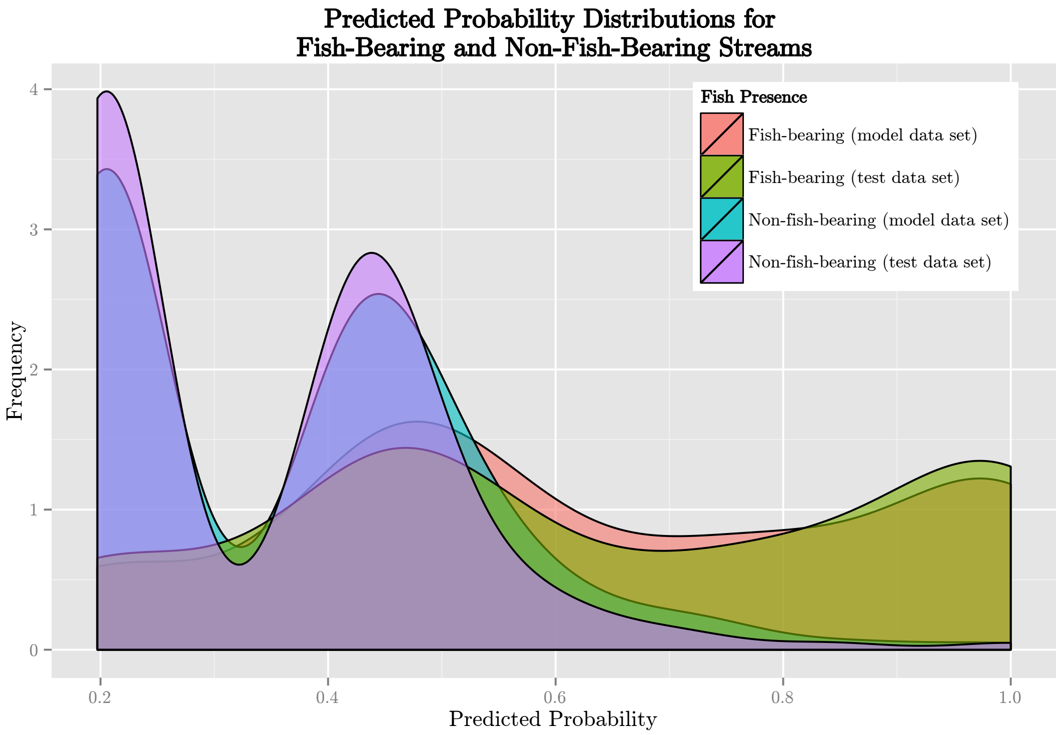


Figure E.14: Overlapping distributions of probability frequencies from model 7 (modelling and testing data sets).

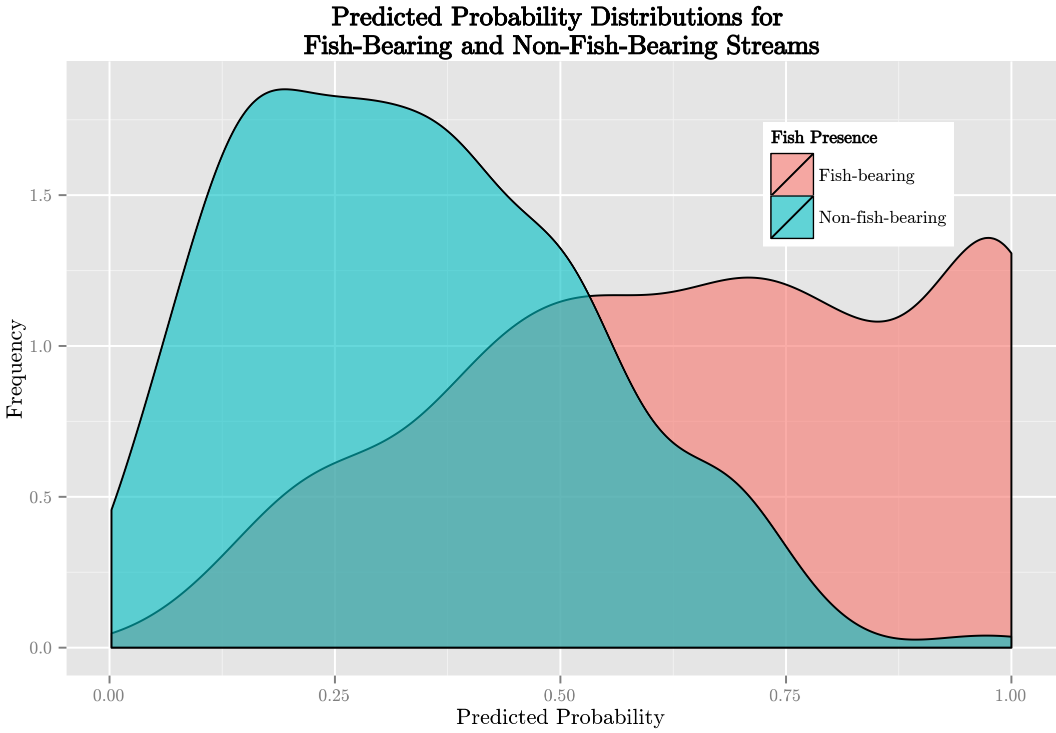


Figure E.15: Overlapping distributions of probability frequencies from model 5c (combined data set).

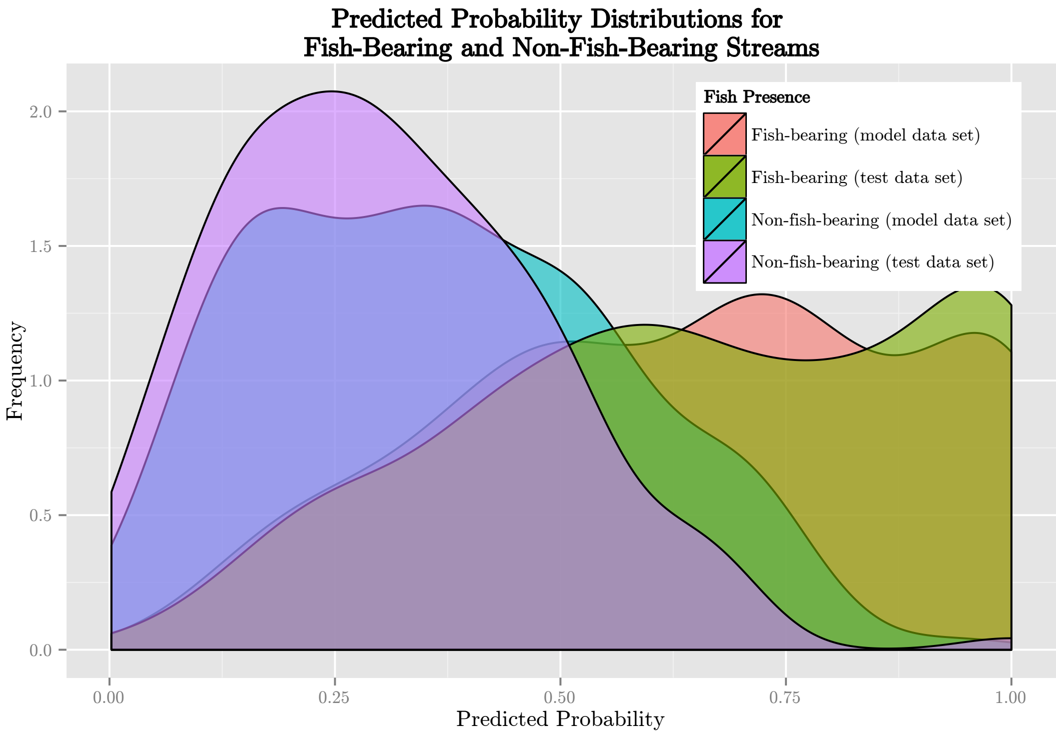


Figure E.16: Overlapping distributions of probability frequencies from model 5c (modelling and testing data sets).