Didymosphenia geminata bloom formation in New Zealand’s rivers

Jonathan McCallum

a thesis submitted for the degree of
Master of Science in Ecology
at the University of Otago, Dunedin,
New Zealand

27 January 2014
Abstract

The diatom *Didymosphenia geminata* (didymo) is recognised as a nuisance bloom-forming species in freshwater systems globally. In New Zealand, didymo has invaded many rivers in the South Island but, to date, has not established in the North Island. Research suggests that didymo bloom formation in rivers is controlled primarily by ambient dissolved reactive phosphorus concentration and flow regime. The purpose of this study was to provide a better understanding of the relationship between streambed stability and didymo bloom formation. In the first part of my thesis, I performed a survey at the reach scale (50 metres along the bank) in forty rivers across the South Island of New Zealand comparing didymo standing crop and cover proportion to the Pfankuch index, a qualitative bed stability assessment tool. Pfankuch values, dissolved reactive phosphorus concentrations and turbidity measurements were compared as predictors of didymo standing crop and cover using an information-theoretic approach. For the nine sites with visible didymo blooms, Pfankuch bed stability was the best predictor (in terms of model likelihoods) of didymo standing crop and cover. These results suggest that streambed stability – evaluated using the Pfankuch index – is an important environmental variable controlling the formation of didymo blooms.

Previous attempts at modelling the distribution of didymo have focused on the potential for didymo cells to survive in waterways rather than the potential for blooms to form. In the second part of my thesis, I modelled the relative suitability of river segments in the New Zealand river network (mean length = 672m, SD=642m) for didymo bloom formation. I used three distinct distribution modelling algorithms (logistic regression, boosted regression trees and Maxent) and compared the results in terms of model structure, model performance and model predictions. I found that the choice of distribution modelling algorithm had a minor influence on the model predictions at a regional scale but a much greater influence at
the river segment scale. The models produced reliable predictions for the South Island and indicated that highly susceptible river segments are concentrated in the Otago, Canterbury and West Coast regions. Additionally, the predictions suggest that a majority of river segments in the North Island are likely to remain free of didymo blooms. Any didymo management strategy should therefore consider the suitability of rivers for didymo bloom formation.
Acknowledgements

I wish to sincerely thank those who assisted, supported and encouraged me during the past year:

- Christoph Matthaei and Gerry Closs for providing guidance and advice. Christoph, you have been a strong mentor and a friend. I benefited greatly from your keen insights and editorial flair (all mistakes remain my own). Thank you Gerry for helping me to attend the New Zealand Freshwater Sciences Society conference in both 2012 and 2013, a gold-mine for intellectual discussion.

- Cathy Kilroy, Phoebe Zarnetske and Vanessa Hammond for sharing your individually vast ecological knowledge. Your input shaped the direction of this work.

- Nicky McHugh, Murray McKenzie and Julie Clark for technical assistance.

- Katha Lange, Andi Bruder, Romana Salis, Jay Piggott and Javad Ramezani for sustaining a supportive lab group.

- Tsukushi Kamiya, Katie O’Dwyer and Boris Berckhout for passionate debates, coffee breaks and other interesting distractions. I count you as close friends.

- Sophie Gibson for the early-morning writing sessions.

- Post-graduate students in the Department of Zoology. I wish you all the best for your own theses.

- My parents – Janet and Michael McCallum – for your patience. I love you both.

- The wonderful Katie Goode for everything.

Thank you all!
Contents

1 General Introduction ............................................ 1
   1.1 The rise of biological invasions ............................. 1
   1.2 Biological invasions in aquatic ecosystems ................. 2
   1.3 Impacts of biological invaders in New Zealand’s freshwaters .... 3
   1.4 Didymo: a recent invader in New Zealand ................... 5
   1.5 Thesis structure ............................................. 7

2 Streambed stability and *Didymosphenia geminata* bloom formation 8
   2.1 Introduction ................................................ 8
   2.2 Methods .................................................... 11
   2.3 Results ..................................................... 15
   2.4 Discussion ................................................ 21

3 Modelling the distribution of *Didymosphenia geminata* blooms in New Zealand 25
   3.1 Introduction ................................................ 25
   3.2 Methods .................................................... 31
   3.3 Results ..................................................... 41
   3.4 Discussion ................................................ 52

4 General Discussion ............................................. 57
   4.1 Study Limitations .......................................... 57
   4.2 Further Research Needs ..................................... 60
   4.3 Concluding note ............................................ 61

References ...................................................... 62

A Didymo survey data ........................................... 73

B Pfankuch index form ........................................... 76

C Violin plots ................................................... 78
Chapter 1

General Introduction

‘If we look far enough ahead, the eventual state of the biological world will become not more complex but simpler – and poorer’ (Elton, 1958)

1.1 The rise of biological invasions

In the opening paragraph of *The Ecology of Invasions by Plants and Animals*, Elton (1958) introduced the idea of an ecological explosion. He drew similarities between an ecological explosion and the spread of the influenza virus at the end of the First World War: there is a point of introduction followed by relatively rapid spread and undeniably disastrous consequences. Today, the phenomenon of one species rapidly becoming dominant in a new area is known as a biological invasion (Valéry et al., 2008).

The movement of humans between all regions of the world has resulted in countless species introductions. Somewhat fortunately, only a fraction of all introductions have led to biological invasions with measurable ecosystem impacts. For example, humans have introduced at least 139 species to the Laurentian Great Lakes in North America (Mills et al., 1993). Not all of these species have established viable populations – the European flounder *Platichthys flesus* and Chinese mitten crab *Eriocheir sinensis*, for instance, have been discovered multiple times but are incapable of reproducing in freshwater (Ricciardi, 2006). Fewer than 10% of the species introduced to the Great Lakes have had measurable ecosystem impacts (Mills et al., 1993).

Human-assisted invasions rose in frequency with the colonisation of new land by people. The increasing connectivity of the world allowed long-distance dispersal events to occur with much greater frequency than the natural rate (Ricciardi, 2007). Acclimatisation societies arose in many countries in the 19th century with the deliberate
intention of homogenising the floras and faunas of distant countries (Nentwig, 2007), further increasing the frequency of biological invasions. There is no precise answer to the question of how many invasions are human assisted, but the number is high enough to warrant much concern.

A note on terminology: In the biological invasion literature, there are several definitions for the terms ‘invader’ and ‘invasive’. For the purpose of this thesis, any organism at the centre of a biological invasion is an invader. Where possible, I have avoided using the term ‘invasive’ because it has been criticised as being too subjective and imprecise for ecology (Colautti and MacIsaac, 2004). It is also worth noting that I use the phrase ‘introduced species’ to describe species that are found outside their native range due to human-mediated dispersal only (following Pysek et al., 2004) and the phrase ‘non-indigenous species’ to describe any species found outside its native range, no matter the circumstances.

1.2 Biological invasions in aquatic ecosystems

The first step in any biological invasion is a dispersal event and, for aquatic ecosystems, that event is often linked to international shipping. Ships move cargo halfway across the planet in just a few weeks. Many organisms can survive in shipping ballast water for this time. Indeed, it is estimated that a single ship may carry in the order of $10^8$ invertebrate individuals (Duggan et al., 2005) – a number of equal magnitude to the human population of Japan. A prime example of a species that was dispersed by international shipping is the zebra mussel. The mussels were transported in shipping ballast from Eurasia to the Great Lakes in the late eighties (Mills et al., 1993). Zebra mussels compete with native amphipod species for seston, leading to declines in amphipod food-resources for the fish community (Vanderploeg et al., 2002). Their fast reproductive rate allows rapid accumulation of biomass on submerged surfaces, causing mortality in native mussel species (Haag et al., 1993) as well as additional costs for marine industries (Connelly et al., 2007).

There are other avenues of human-mediated dispersal in aquatic ecosystems apart from shipping. The aquarium trade, for instance, has been implicated in spreading numerous aquatic species across the world (Padilla and Williams, 2004). Aquarium owners are thought to have (unintentionally) caused hundreds of aquatic plant species introductions by disposing unwanted pets and the contents of aquaria into waterbodies. The International Union for the Conservation of Nature (IUCN) maintains a list of the
100 worst invasive species and one third of the species are, or were, plants in the aquarium trade (Padilla and Williams, 2004).

The impacts of a single biological invasion on an aquatic ecosystem are difficult to predict because they depend on the properties of the invader, the abundance of the invader and the composition of the community being invaded (Lodge, 1993). Fortunately, the majority of species introductions never result in a biological invasion because the abundance of the introduced species remains low or the interactions between it and the in situ community are weak (Mills et al., 1993).

Despite most species introductions having little impact on aquatic ecosystems, the remaining few are responsible for extensive damage. Three examples are the introduced marine alga, Caulerpa taxifolia which has reduced more than 13,000 hectares of seabed in the Mediterranean Sea to a monoculture (Meinesz et al., 2001); the water flea Cer-copagis pengoi which voraciously predates on fish larvae and may compete with other planktivores in the Great Lakes (Laxson et al., 2003) and the water hyacinth Eichhornia crassipes which has invaded freshwaters in over 50 countries, modifying water chemistry and reducing phytoplankton production (Villamagna and Murphy, 2010).

1.3 Impacts of biological invaders in New Zealand’s freshwaters

New Zealand’s freshwaters contain many non-indigenous species. There are at least 50 non-indigenous aquatic macrophytes (Coffey and Clayton, 1988; Champion and Clayton, 2001) and 21 non-indigenous fish species (Dean, 2001; McDowall, 2010) – accounting for more than half of the macrophyte flora and a third of the fish fauna, respectively. Of the non-indigenous fish fauna, eight species have highly-restricted distributions and two species may not establish viable populations, but the remaining eleven (including catfish, koi carp, pike, tilapia and several salmonids) have invaded or have the potential to invade lake or river habitat (Dean, 2001). Other species introduced to New Zealand’s freshwaters include numerous invertebrates (Nentwig, 2007) and an unknown number of microorganisms.

Although it is likely that most species introduced to New Zealand fail to establish a viable population, the few that do successfully invade demand our attention. Two New Zealand invaders with well-documented impacts are brown trout, a species introduced for sports fishing, and the aquatic macrophyte Lagarosiphon major.

The brown trout (Salmo trutta) has a long history in New Zealand. Acclimatisation societies introduced Tasmanian troutlings to lakes, rivers and streams from the 1860s
(Flecker and Townsend, 1994). The introduction programme continues today with Fish and Game councils stocking fisheries in the North and South Islands (Townsend, 1996). Although trout stocking encourages recreational use of New Zealand’s waterways, there are costs, and these costs are largely realised through ecological impacts.

Brown trout are highly aggressive and territorial (Hearn, 1987). Their competitive behaviour has contributed to the range contraction of numerous native fish, primarily in the family Galaxiidae. Galaxiid numbers are also reduced because trout eat galaxiids (Crowl et al., 1992). So extreme has their range contraction been, that several galaxiid species are now found only in headwaters above steep waterfalls (Townsend, 1996). Comparatively moderate impacts of brown trout introductions on stream invertebrates have also been observed. For instance, mayfly grazing shifts from day to night (Mcintosh and Townsend, 1995) and insect densities decline (Flecker and Townsend, 1994).

Nine decades after the introduction of brown trout, *Lagarosiphon major* was first detected in New Zealand lakes (in 1950; Howard-Williams and Davies, 1988). Originating from Southern Africa, *Lagarosiphon* has become the dominant macrophyte in New Zealand clear-water lakes, including Lake Taupo, Lake Rotorua (Schwarz and Howard-Williams, 1993) and Lake Wanaka (Kelly and Hawes, 2005). It is capable of vigorous growth with a tall habit such that *Lagarosiphon* forests up to several metres deep are common in these lakes (Schwarz and Howard-Williams, 1993).

The impacts of *Lagarosiphon* on New Zealand’s lake ecosystems are well-known. The thin stems with radially arranged leaves can quickly outcompete native macrophytes for light resources (Howard-Williams and Davies, 1988). The change in physical structure of weed beds following *Lagarosiphon* invasion alters invertebrate communities, increasing the relative abundance of chironomids and caddisflies but reducing the abundance of oligochaetes and nematodes (Kelly and Hawes, 2005). On a more positive note, there is evidence that *Lagarosiphon* beds provide suitable habitat for native fish (Bickel and Closs, 2008), and they appear to be a food source for swans (Howard-Williams and Davies, 1988). However, *Lagarosiphon* forests impact humans by degrading the visual appeal of lakes for water users, impeding the launching of boats, becoming tangled in fishing gear and generally reducing the recreational value of New Zealand’s lakes (McGregor and Gourlay, 2002). Due to its biological success and weediness, *Lagarosiphon* was classified as one of the eight worst aquatic weeds in New Zealand (Champion and Clayton, 2001).

I have used these two examples to illustrate the point that biological invasions have affected New Zealand’s freshwaters in many ways. Not only have brown trout reduced habitat for native galaxiidae species dramatically; and have *Lagarosiphon* forests grown
to suffocate lakes, but both species will continue to do so for the foreseeable future. Clearly, these impacts are worthy of study and efforts to control them.

### 1.4 Didymo: a recent invader in New Zealand

New Zealand has a long history of invasions in its freshwater ecosystems, as described in the previous sections. Recently, the invasion of New Zealand by the diatom *Didymosphenia geminata* (didymo) has received much attention due to the extensive blooms that it forms in rivers.

Based on fossil records, didymo is thought to be native to Asia, Europe and parts of North America (Blanco and Ector, 2009). For most of its history in these continents, didymo was considered a rare taxon (Whitton *et al.*, 2009). So it was surprising that, in 2004, didymo grew to 100% cover of parts of the East Boulder river in Montana (Whitton *et al.*, 2009). Since then it has spread in a way that has been likened to a pandemic (Kirkwood *et al.*, 2009) forming nuisance blooms within its native range (Blanco and Ector, 2009) and invading the Southern Hemisphere countries of New Zealand, Chile and Argentina (Kilroy and Unwin, 2011; Segura, 2011; Reid and Torres, 2014).

In late 2004 didymo was detected in the Lower Waiau river in Southland, New Zealand, during routine monitoring work by the National Institute of Water and Atmospheric Research (Kilroy, 2004). Over the next four years it spread north, reaching the top of the South Island (MAF Biosecurity NZ, 2012). This range expansion is most likely human-assisted by the transport of didymo cells in fishing or kayaking equipment (Kilroy and Unwin, 2011). It may be possible for didymo to establish in the North Island but, to date, it has not yet been detected there. A campaign to reduce the spread of didymo was set up from its earliest detection by government agencies.

Under favourable conditions, didymo forms conspicuous mats on the riverbed. The mats are complex biofilms primarily containing didymo cells and extracellular polysaccharide (EPS) stalk material produced by the cells (Kilroy and Bothwell, 2011). Other organisms are found attached to the EPS or associated with the mats, such as cyanobacteria, small-celled diatoms and other epiphytic algae (Whitton *et al.*, 2009). The thick (up to 20cm as reported by Spaulding and Elwell, 2007) interwoven structure of the mats are also likely to contain trapped inorganic material and fine particulate organic matter. Didymo mats have been documented completely covering several kilometres of a riverbed (Kilroy *et al.*, 2009). Mats of this magnitude are considered to be nuisance blooms (didymo blooms) to recreational water users – fisherman, for example, report that their fishing gear becomes tangled in clumps of mat material. The mats also
are visually unappealing, looking alarmingly like floating pieces of toilet paper when detached.

Studies have revealed ecological impacts of didymo blooms on river ecosystems. Perhaps an obvious impact is that, where didymo blooms occur, the algal community becomes dominated by didymo cells and associated EPS (Kilroy et al., 2009). Algal community biomass and algal species richness may increase in some areas because of the numerous epiphytes associated with the mats (Whitton et al., 2009).

These changes in algal community biomass and composition have a flow-on impact for the fauna. Several studies have shown that didymo blooms increase invertebrate densities (Kilroy et al., 2009; Gillis and Chalifour, 2009; James et al., 2010). There are, however, conflicting results of the impact of blooms on invertebrate species richness and invertebrate community diversity. In a survey of five sites, across eleven sampling occasions in Rapid Creek, South Dakota, Larson (2007) detected a moderate negative relationship between invertebrate species richness and didymo biomass but no relationship with Simpson diversity. In contrast, Kilroy et al. (2009) found a moderate positive relationship between species richness and didymo blooms in surveys of the Mararoa (6 sites) and Oreti (5 sites) rivers in New Zealand. These conflicting results and relatively small number of sites in the two studies suggest that it is too early to draw a conclusion about the relationship between invertebrate diversity, or invertebrate biomass and didymo blooms. It is also unclear to what extent didymo blooms affect the fish community of rivers.

It is clear, however, that the composition of the invertebrate community is altered by didymo blooms. The community declines in EPT taxa (Ephemeroptera, Plecoptera and Trichoptera; Larson, 2007; James et al., 2010) and there are increases in chironomid and oligochaete abundances (Kilroy et al., 2009). These changes represent a shift in the invertebrate community towards smaller, sediment-loving fauna.

The bloom-forming growth habit of didymo also has important implications for human activities linked to streams and rivers in New Zealand. For example, clumps of algal material – removed from the benthos in high flows – can float downstream where they have the potential to clog the intakes of irrigation equipment and foul the races and canals of hydro-electric power stations (Spaulding and Elwell, 2007). In addition, the detached mats can accumulate on the banks and, as with *Lagarosiphon*, reduce the recreational value of New Zealand’s freshwaters.
1.5 Thesis structure

Considering the impacts of didymo blooms, two pressing questions are ‘What are the conditions required for didymo blooms to form?’ and ‘Where are those conditions found in New Zealand?’. A clear answer to these questions will be helpful when reviewing didymo management plans. In this thesis I help to answer these questions and contribute to our understanding of didymo bloom formation, with a focus on New Zealand.

In Chapter 2, I examine the formation of didymo blooms in relation to streambed stability – a physical characteristic of rivers. I examine whether streambed stability is a useful predictor of didymo standing crop and percentage cover (two variables relating to didymo bloom extent) using a field survey.

In Chapter 3, I model the potential distribution of didymo blooms in New Zealand. I use three species distribution modelling algorithms – logistic regression, boosted regression trees and Maxent – to combine didymo bloom location data with information on the environmental conditions found in New Zealand’s rivers. The result is a set of prediction maps indicating the relative suitability of river segments for didymo bloom formation. By using three species distribution modelling algorithms, the predictions from each model and the performance of each model could be compared.

In the final chapter, I address key limitations of my work and I include several recommendations for future work.
Chapter 2

Streambed stability and

Didymosphenia geminata bloom formation

2.1 Introduction

Biological invasions in aquatic ecosystems are of immediate global concern. The impacts of these invasions range from the alteration of food webs to the driving of endemic species towards extinction (Simon and Townsend, 2003; Mooney and Cleland, 2001). If the impacts of any particular biological invasion are to be reduced, it is essential that the ecology of the invading organism be well understood.

Didymosphenia geminata (didymo) is a periphytic stalk-forming diatom which has gained notoriety for forming dense mats of stalk material (blooms) on riverbeds and lakeshores (Spaulding and Elwell, 2007). Paleontological observations indicate that didymo is native to the northern regions of Europe, Asia and North America (Blanco and Ector, 2009). Didymo was detected in New Zealand in 2004 and subsequently spread across the South Island (Kilroy and Unwin, 2011). In 2010, didymo was observed in the Patagonian rivers of South America (Segura, 2011). In recent decades, persistent didymo blooms have been documented in the species’ native range (Kawecka and Sanecki, 2003; Bhatt et al., 2007; Bothwell et al., 2009; Kirkwood et al., 2009) and in the invaded range (Patagonia and New Zealand’s South Island; Segura, 2011; Kilroy and Unwin, 2011).
Invasion, Growth and Removal

Didymo can produce large amounts of stalk material in waters with very low nutrient concentrations (oligotrophic) while experiencing strong shear forces from the water passing over the streambed (Spaulding and Elwell, 2007). This phenomenon – high biomass under oligotrophic conditions – has been termed the didymo paradox (Cullis et al., 2012).

In the conceptual model proposed by Cullis et al. (2012), didymo mat establishment can be broken into four stages:

- **Invasion** facilitated by a vector, often human-mediated (Kilroy and Unwin, 2011)
- **Growth** controlled by numerous environmental variables
- **Removal** by shear stress, abrasion or bed disturbance (scouring)
- **Recovery** after a disturbance

Several studies of didymo’s ecology have focused on the growth stage, revealing the importance of phosphorus in controlling blooms. For instance, standing crop and percentage cover of didymo were both negatively related to phosphorus concentration in a survey of 19 rivers in the South Island (Kilroy and Bothwell, 2012). This contrasts with the growth patterns of most benthic algae which tend to reach higher biomass at higher nutrient concentrations (Biggs and Close, 1989).

The observation that didymo proliferates in oligotrophic conditions has encouraged a closer inspection of the mechanism involved (Cullis et al., 2012). When phosphorus is low, the photosynthate derived from the cells is primarily exuded as extracellular polysaccharide (Kilroy and Bothwell, 2012). This stalk material may increase nutrient availability by pushing the cells further into the flow, allowing access to more nutrients. Under mesotrophic or eutrophic conditions, however, cell division dominates over stalk production resulting in lower biomass overall (Bothwell and Kilroy, 2011).

Another important factor controlling didymo blooms is flow stability. Evidence for this has arisen from examining rivers with dams. Sites downstream of dams experience more stable flows and have a higher density of didymo cells than those upstream (Kirkwood et al., 2009). Hammond (2013) found a similar result after measuring ash-free dry mass in 21 didymo-positive waterways.
**Bed disturbance and Periphyton**

A disturbance is the application of forces to an ecosystem causing some form of damage, be it the death or displacement of individuals or the destruction or depletion of resources (Lake, 2000). In lotic ecosystems, floods are a prevalent, and widely studied, form of disturbance (Biggs *et al*., 1999; Lake, 2000). Floods involve an increase in water velocity and depth compared to a reference condition. For benthic biota, a commonly used reference is the median flow over one year and an appropriate threshold for flood disturbance is three times this reference value (Clausen and Biggs, 1997). Floods often cause the suspension of fine particles and movement of the substratum (Biggs *et al*., 1999). This movement, or bed disturbance, is an important organising factor for periphyton communities (Biggs, 1995).

Species differ in their response to disturbance (Lytle and Poff, 2004). Some periphyton species have adapted attachment methods which allow them to tolerate higher water velocities without experiencing shear removal (Biggs *et al*., 1998). Other periphyton bind the substratum together, increasing the magnitude of a flood required to cause bed movement.

During the majority of floods, not all patches of a streambed move. Rather, streams often exhibit small scale patterns of sediment deposition, scour or stability (lack of movement) during flood disturbance (Matthaei *et al*., 1999b; Matthaei and Townsend, 2000). A streambed’s local disturbance history, incorporating the spatial and temporal pattern of substratum movement, has been shown to have a strong influence on the small-scale distribution of periphyton (Matthaei *et al*., 2003). At the reach scale too, flood-induced physical disturbance of the streambed is a primary habitat variable altering periphyton community composition and biomass (Biggs *et al*., 1998).

Two techniques used by researchers to estimate streambed stability are the tracer stone and scour chain methods. The tracer stone method involves marking a set number of particles with paint or radio frequency ID tags (Death and Zimmermann, 2005; Schwendel *et al*., 2011). The particles are removed from the stream, marked and then carefully repositioned. Particle position is recorded before and after a bed-moving flood and, later, the proportion of particles which moved and the distance they moved may be incorporated into a multivariate index of bed stability (eg. Death and Winterbourn, 1994).

The scour chain method involves installing metal chains vertically into the streambed (Schwendel *et al*., 2010). Researchers are then able to record the three-dimensional patterns of scour and deposition during a bed-moving flood. The scour chain method has been used to link local disturbance history with the small-scale distribution of
periphyton (Matthaei et al., 2003) and stream fauna (Matthaei and Townsend, 2000; Effenberger et al., 2006, 2008).

Both the tracer stone and scour chain methods are laborious. In contrast, the Pfankuch bed stability index (Pfankuch, 1975) is a checklist style approach to estimating streambed stability that can be completed in 15 minutes at a given field site (Schwendel et al., 2011). The Pfankuch index is intended to be applied at the reach scale and it encompasses several different aspects of streambed stability including patterns of scour and deposition, particle consolidation plus substratum shape and size (Pfankuch, 1975).

**Objectives and Hypotheses**

Adding to research on didymo growth, cited above, the first objective of this research was to examine the relationship between flood-induced physical disturbance of the streambed and didymo bloom formation. I expected low didymo standing crop and cover in unstable reaches (higher Pfankuch values) and high didymo standing crop and cover in very stable reaches (low Pfankuch values).

The second objective was to compare Pfankuch value, dissolved-reactive phosphorus concentration and turbidity as predictors of didymo standing crop and cover. I expected that Pfankuch value would be a better predictor of standing crop and cover, in terms of model likelihood, than dissolved reactive phosphorus or turbidity.

**2.2 Methods**

**Study Sites**

Forty river sites (Figure 2.1) were surveyed in the South Island of New Zealand between latitudes 42.4 and 46.2 degrees south during stable flows. For this study, stable flows were defined as those below three times the median value over the previous six months, a commonly used threshold to characterise flood frequency in New Zealand rivers (Clausen and Biggs, 1997). Waterways were only surveyed if they had a history of detection of didymo cells recorded in the Didymo Samples Database (MAF Biosecurity NZ, 2012). One site was placed in each waterway within a reach of Strahler stream order four to eight. Few, if any, reaches below Strahler stream order four have a history of detection of didymo cells in New Zealand (MAF Biosecurity NZ, 2012). Stream order was determined using the NIWA Water Resources Explorer (WRENZ, 2012). Study sites were placed by marking 50m along the bank. Where possible, sites
included both riffle and run habitats, were within 200m of a road and had sparse, if any, canopy cover to reduce the influence of light availability on didymo standing crop. In addition, each site was chosen to have comparable water depths and current velocities.

**Nutrient and Periphyton Samples**

Water samples were collected at the downstream end of reaches using 50ml polypropylene tubes before walking into other parts of the reach. Two samples were filtered through 25mm micro-glass fibre paper (Munktell, Germany) on-site after rinsing the polypropylene tubes in stream water. Duplicate, unfiltered samples were collected at the same time. All sample tubes were transported to the lab chilled, and were then frozen at -20°C. Nitrogen and phosphorus analyses were performed on thawed filtered samples with a FIAstar 5000 flow injection analyser (Foss, Denmark). Turbidity measurements were taken on thawed unfiltered samples using a laboratory benchtop turbidimeter.

Qualitative periphyton samples were collected by scraping the top surface of three randomly chosen cobbles into a sample container. Approximately 1ml of each sample was viewed under a compound microscope at 400x magnification to visually confirm the presence of *D. geminata* cells.

**Standing Crop and Cover**

At each site, cover of didymo mats was estimated using a variation of the point quadrat method (as used by Leonard and Clark, 1993). First, a 6m chain was placed on the benthos at the downstream end of the study reach. Second, 20 markings on the chain – at 25cm intervals – were used to score the presence or absence of didymo mats. Presence was defined as any visible didymo colony directly below the marking. Concurrently, mat depth was measured using a small ruler. This procedure was repeated every 10 metres, moving upstream, along the 50m reach (6 times in total to give 120 points).

From these data, didymo cover was calculated by finding the proportion of points with didymo mats. Didymo cover, as a percentage, was converted to didymo standing crop index (SCI) by multiplying by mean mat thickness:

\[
SCI = \text{cover}\% \times \text{mean}(\text{thickness})
\]

Standing crop provides a reliable estimate of didymo biomass (Kilroy and Bothwell, 2012) and a standing crop value of approximately 220 corresponds with 35gm⁻² AFDM, the threshold for nuisance periphyton in New Zealand.
Pfankuch bed stability

Bed stability was estimated using the bottom component of the Pfankuch bed stability index (Appendix B). The bottom component was used because it relates more strongly to quantitative stability measures than the full index (Death and Winterbourn, 1994). Two observers filled out the Pfankuch index evaluation form separately after walking the full length of the site (recommended by Pfankuch, 1975). One question from the index, regarding ‘clinging aquatic vegetation’ was excluded because it involves an assessment of algal cover, a potential source of autocorrelation. The index ranged from 14 (very stable) to 56 (very unstable).

In addition to the Pfankuch index, the particle composition of the river bed was estimated by assigning a percentage to each of the following seven size classes: bedrock, boulders (>250mm in length), large cobbles (120-250mm), small cobbles (60-120mm), gravel (12-60mm), sand and silt. The same categories were used by Kilroy and Bothwell (2012). Temperature measurements were recorded using a handheld multiparameter instrument (YSI Professional Plus). Surface current velocity was measured by timing the movement of a tennis ball over 10m, repeating five times and calculating the mean.

Data Analysis

Principal component analysis was conducted on 11 standardised environmental variables to examine the strength of the correlations between them. Principal component analysis was also carried out on the variables within the Pfankuch index to determine whether sites with visible didymo mats clustered together. Both analyses were performed using the vegan R package (Oksanen et al., 2013) following Borcard et al. (2011).

An information-theoretic approach was used to assess the candidate hypotheses (Table 2.1), each with a single response and a single predictor variable. The response variables were the percentage cover of didymo and didymo standing crop. The predictor variables were Pfankuch bed stability, dissolved reactive phosphorus and turbidity. Normal linear regression was used to fit each predictor to standing crop index and percentage cover. Prior to model fitting, standing crop was log-transformed and percentage cover was arcsine square-root transformed to make the data more normal (Whitlock and Schluter, 2009).

The models shown in Table 2.1 were fitted to two datasets, one containing data from all sites, the other containing only the sites with positive values for didymo standing crop. Multiple regression models – incorporating several predictors – were considered but ultimately not fitted due to the small number of non-zero response measurements.
Table 2.1: Candidate models relating Pfankuch bed stability (Pfankuch), dissolved reactive phosphorus (DRP) and turbidity to didymo standing crop and cover proportion.

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Model</th>
<th>Predictor</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing crop</td>
<td>$H_1$</td>
<td>Intercept-only</td>
<td>$Y_1 = \beta_0$</td>
</tr>
<tr>
<td>(log[Y+1] transformed)</td>
<td>$H_2$</td>
<td>Pfankuch ($X_1$)</td>
<td>$Y_1 = \beta_0 + \beta_1 X_1$</td>
</tr>
<tr>
<td></td>
<td>$H_3$</td>
<td>DRP ($X_2$)</td>
<td>$Y_1 = \beta_0 + \beta_1 X_2$</td>
</tr>
<tr>
<td></td>
<td>$H_4$</td>
<td>Turbidity ($X_3$)</td>
<td>$Y_1 = \beta_0 + \beta_1 X_3$</td>
</tr>
<tr>
<td>Cover proportion</td>
<td>$H_5$</td>
<td>Intercept-only</td>
<td>$Y_2 = \beta_0$</td>
</tr>
<tr>
<td>(arcsine square root transformed)</td>
<td>$H_6$</td>
<td>Pfankuch ($X_1$)</td>
<td>$Y_2 = \beta_0 + \beta_1 X_1$</td>
</tr>
<tr>
<td></td>
<td>$H_7$</td>
<td>DRP ($X_2$)</td>
<td>$Y_2 = \beta_0 + \beta_1 X_2$</td>
</tr>
<tr>
<td></td>
<td>$H_8$</td>
<td>Turbidity ($X_3$)</td>
<td>$Y_2 = \beta_0 + \beta_1 X_3$</td>
</tr>
</tbody>
</table>

Following model fitting, the corrected Akaike Information Criterion, AICc, was calculated for each model. AICc applies a penalty to models with more parameters and awards models which fit data more closely, in terms of log-likelihoods (Burnham et al., 2011):

$$AICc = -2 \log(L(\hat{\theta})) + \frac{2k(k + 1)}{n-k-1}$$

Here, $\log(L(\hat{\theta}))$ is the log-likelihood of the model, $k$ is the number of parameters in the model and $n$ is the sample size. Models with the lowest AICc have the best balance between model fit and model complexity (Anderson, 2008). Prior to comparing model likelihoods, the difference in AICc ($\Delta AICc$) was calculated. Model sets with all $\Delta AICc$ values less than 7 were deemed to have no ‘best’ model (as per Anderson, 2008; Burnham et al., 2011).

The evidence ratio (ER) and coefficient of determination ($r^2$) were also calculated. The evidence ratio is the likelihood of one model relative to another (Symonds and Moussalli, 2011). It was calculated from $\Delta AICc$ of the two models, with the lower value in the numerator ($\Delta_l$) and the higher value in the denominator ($\Delta_h$) of the following fraction:

$$ER = \frac{exp(-\frac{1}{2} \Delta_l)}{exp(-\frac{1}{2} \Delta_h)}$$

I used the coefficient of determination ($r^2$) as a measure of effect size. To account for the bias in $r^2$ at small sample sizes (Nakagawa and Cuthill, 2007), 95% confidence intervals were included and interpreted. All analyses were conducted in R (R Core Team, 2013).
2.3 Results

Figure 2.1: Survey sites in the South Island of New Zealand. Each site was visited once in November 2012. Abbreviations of the waterway name are shown next to each point.

Forty sites were surveyed in a period of 12 days during November 2012. Didymo mats were visible at nine of the 40 sites and ranged from scattered button-sized colonies (standing crop index = 12.5) to mats greater than 20cm deep, covering almost the entire riverbed (standing crop index > 2800). Flow monitoring graphs made public by regional councils provided hourly flow information from monitoring stations near 24 sites in this survey. Of these, 20 stations recorded flows less than three times the
median flow in the previous month (Environment Canterbury, 2012; Otago Regional Council, 2012; Environment Southland, 2013). Four monitoring stations recorded flood flows in the same period (Table 2.2). Visible didymo mats were not found at the four sites nearest to these monitoring stations. Considering their magnitude, it is highly likely these four events were sufficient to remove didymo mats from these sites.

Table 2.2: Flow peaks recorded between 19th October and 18th November 2012. Distance is the channel distance of the monitoring station upstream (us) or downstream (ds) of the nearest site. ‘Median’ is the one-year median flow recorded at the monitoring station. Flow data from Environment Southland (2013).

<table>
<thead>
<tr>
<th>Monitoring Station</th>
<th>Nearest Site</th>
<th>Distance (km)</th>
<th>Median (cumecs)</th>
<th>Peak (cumecs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waiau at Tuatapere</td>
<td>LWAI</td>
<td>15 ds</td>
<td>65</td>
<td>800</td>
</tr>
<tr>
<td>Mararoa at The Cliffs</td>
<td>MARA</td>
<td>1 ds</td>
<td>19</td>
<td>150</td>
</tr>
<tr>
<td>Aparima at Dunrobin</td>
<td>APA</td>
<td>22 us</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Mataura at Seaward Downs</td>
<td>MAT</td>
<td>31 ds</td>
<td>70</td>
<td>460</td>
</tr>
</tbody>
</table>

The mean, standard deviation and range of the environmental variables measured across the 40 sites are shown in Table 2.3. Pfankuch bed stability scores covered approximately 85% of the full range of possible values.

Table 2.3: Mean, standard deviation (St Dev) and range of environmental variables for the 40 sites surveyed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>St Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average depth (cm)</td>
<td>34</td>
<td>10</td>
<td>18</td>
<td>60</td>
</tr>
<tr>
<td>Surface current velocity (cms⁻¹)</td>
<td>81</td>
<td>23</td>
<td>32</td>
<td>134</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>13.6</td>
<td>2.8</td>
<td>6.8</td>
<td>18.6</td>
</tr>
<tr>
<td>Pfankuch bed stability</td>
<td>42</td>
<td>11</td>
<td>18</td>
<td>54</td>
</tr>
<tr>
<td>Dissolved reactive phosphorus (µgL⁻¹)</td>
<td>2.7</td>
<td>1.8</td>
<td>0.5</td>
<td>8.5</td>
</tr>
<tr>
<td>Nitrate-nitrite nitrogen (µgL⁻¹)</td>
<td>90</td>
<td>160</td>
<td>0</td>
<td>865</td>
</tr>
<tr>
<td>Turbidity (NTU)</td>
<td>0.9</td>
<td>1.2</td>
<td>0.2</td>
<td>5.6</td>
</tr>
</tbody>
</table>

For the principal component analysis of the environmental variables measured in this study, the first two principal components explained 42% of the variation in the data with component one and component two accounting for 22% and 20%, respectively. In the correlation plot (Figure 2.2), a group of three predictors - Pfankuch bed stability
index (Pfan), concentration of dissolved reactive phosphorus (DRP) and nitrogen concentration (NN) - were on the opposing side of standing crop index (SCI) and didymo cover. This ordination structure suggests the three predictors are positively correlated with each other and negatively correlated with standing crop and cover.

Figure 2.2: Principal component analysis correlation plot of environmental variables for all 40 sites. The angles between variables indicate the strength of the correlation between them. Sharply acute angles indicate positive correlations and angles close to 180 degrees indicate negative correlations (Legendre and Legendre, 1983). Variable codes are as follows: Depth - mean depth, cover - didymo cover, SCI - didymo standing crop, Lat - site latitude, Long - site longitude, Temp - water temperature, Velocity - surface current velocity, Turb - mean turbidity, Pfan - Pfankuch bed stability score, NN - nitrate nitrogen concentration, DRP - dissolved reactive phosphorus concentration.

In the first two principle components of the Pfankuch index variables, no clear clustering of sites was evident (Figure 2.3). Here, principal component one and two explained 57% and 19% of the variation in the data, respectively.

**Didymo standing crop and cover**

The evidence ratios for positive-site models were two orders of magnitude greater than those for all-site models (Table 2.4). Pfankuch bed stability was the best predictor of didymo standing crop and cover for the trimmed dataset ($\Delta$ AICc for all other models in the set > 7). There was little support for the models containing DRP and Turbidity in
Table 2.4: Model fitting summary for didymo standing crop and cover proportion. Models were fitted to the full dataset (all sites) and, separately, sites having positive values for standing crop and cover (positive sites). Corrected Akaike Information Criterion values (AICc) are shown for comparison within each model set. For each set, the model with the lowest AICc is the reference for \( \Delta \text{AICc} \) and the evidence ratio (ER). The estimates for the coefficient of determination \( (r^2) \) have 95% confidence intervals in brackets.

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Predictor</th>
<th>AICc</th>
<th>( \Delta \text{AICc} )</th>
<th>ER</th>
<th>( r^2 ) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing crop</td>
<td>None</td>
<td>174.9</td>
<td>2.5</td>
<td>3.5</td>
<td>0.11 (0 to 0.35)</td>
</tr>
<tr>
<td></td>
<td>Pfankuch</td>
<td>172.4</td>
<td>0</td>
<td>0.11</td>
<td>0.05 (0 to 0.27)</td>
</tr>
<tr>
<td></td>
<td>DRP</td>
<td>174.9</td>
<td>2.5</td>
<td>3.5</td>
<td>0.03 (0 to 0.22)</td>
</tr>
<tr>
<td></td>
<td>Turb</td>
<td>175.9</td>
<td>3.5</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Standing crop</td>
<td>None</td>
<td>39.9</td>
<td>10.3</td>
<td>170</td>
<td>0.78 (0.16 to 0.96)</td>
</tr>
<tr>
<td></td>
<td>Pfankuch</td>
<td>29.6</td>
<td>0</td>
<td>0.78</td>
<td>0.07 (0 to 0.72)</td>
</tr>
<tr>
<td></td>
<td>DRP</td>
<td>42.6</td>
<td>13.0</td>
<td>670</td>
<td>0.07 (0 to 0.71)</td>
</tr>
<tr>
<td></td>
<td>Turb</td>
<td>42.6</td>
<td>13.0</td>
<td>670</td>
<td></td>
</tr>
<tr>
<td>Cover proportion</td>
<td>None</td>
<td>27.3</td>
<td>3.6</td>
<td>6.0</td>
<td>0.13 (0 to 0.38)</td>
</tr>
<tr>
<td></td>
<td>Pfankuch</td>
<td>23.7</td>
<td>0</td>
<td>0.13</td>
<td>0.04 (0 to 0.25)</td>
</tr>
<tr>
<td></td>
<td>DRP</td>
<td>27.7</td>
<td>4.0</td>
<td>7.4</td>
<td>0.02 (0 to 0.20)</td>
</tr>
<tr>
<td></td>
<td>Turb</td>
<td>28.6</td>
<td>4.9</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Cover proportion</td>
<td>None</td>
<td>12.4</td>
<td>7.4</td>
<td>40</td>
<td>0.70 (0.06 to 0.95)</td>
</tr>
<tr>
<td></td>
<td>Pfankuch</td>
<td>5.0</td>
<td>0</td>
<td>0.70</td>
<td>0.03 (0 to 0.66)</td>
</tr>
<tr>
<td></td>
<td>DRP</td>
<td>15.5</td>
<td>10.5</td>
<td>190</td>
<td>0.15 (0 to 0.77)</td>
</tr>
<tr>
<td></td>
<td>Turb</td>
<td>14.4</td>
<td>9.4</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.3: Sites with visible didymo mats (filled circles) and without visible didymo mats (clear circles) plotted on the first two principal components of the Pfankuch index variables (rock angularity, brightness, consolidation, bottom size distribution and scouring/deposition).

all model sets. Models of standing crop and cover for the full dataset had \( \Delta \text{AICc} < 7 \), therefore, no best model was chosen.

The only models with confidence interval for \( r^2 \) above zero (indicating statistical significance) were the Pfankuch bed stability model of standing crop (positive sites only) and Pfankuch bed stability model of cover proportion (positive sites only). Refer Figure 2.4 for plots of fitted models.
Figure 2.4: Pfankuch bed stability as a predictor of didymo standing crop (log\(Y+1\) transformed) and the proportion of site covered with didymo (cover proportion; arcsine square-root transformed). Normal linear models were fitted to data from all sites \((n = 40; \text{A and C})\) and sites with non-zero standing crop values \((n = 9; \text{B and D})\). The dashed lines define 95% confidence bands.
2.4 Discussion

I surveyed forty waterways across the South Island of New Zealand in November 2012 and found didymo mats in 9 out of 40 sites (22.5%). This is a lower percentage of sites with didymo mats than found by Kilroy and Bothwell (2012) in a similar survey (16 out of 31 sites, or 52%, had no didymo mats).

Several factors may have contributed to the low percentage of sites with didymo mats in my survey. Firstly, sampling may have been conducted prior to bloom formation. In a single-river study of didymo biomass dynamics from April 2008 to May 2010, Hammond (2013) showed that didymo mats can reach bloom proportions after 32 to 78 days of accrual. In my study, four sites had accrual periods shorter than this (refer Table 2.2). Secondly, didymo mats may have been present at some sites but were not detected. The modified point quadrat method used to estimate didymo cover and standing crop failed to detect didymo mats where cover was below approximately 5%. At three sites (ASH, CASS and MAK) for example, small dots of didymo were sighted during the fieldwork (J. McCallum, personal observations) but the sampling method was not sensitive enough to detect them. Finally, there may have been a real ecological reason why didymo mats were not detected such as where the substrate is unsuitable for attachment (Bergey et al., 2009). Even though I relied on cell presence records in the Didymo Samples Database (DSDB) to choose the survey sites, some sites may not be suitable for didymo mat establishment. Cells detected at a particular site and recorded in the DSDB may have drifted kilometres downstream from their source. By repeating the biomass assessment in several seasons, the reason for didymo mat absence may become clearer. As multi-seasonal sampling was not feasible for this study, the sources of the zeros in the dataset for standing crop and cover remain unclear.

When the sources of zeros in a dataset are unclear, the zeros are typically excluded and only the sites with positive values are modelled (Martin et al., 2005). I used this approach combined with an analysis of the full, zero-inflated dataset. By fitting models to both datasets, the effect of removing zeros on the Δ AICc values and thus the model ordering could be interpreted and compared. These results are discussed in the next section.

Streambed stability and didymo

Flood-induced physical disturbance of the streambed alters periphyton community composition and biomass (Biggs et al., 1998). Relatively stable reaches have milder physical disturbance regimes in terms of the intensity, frequency, return interval, magnitude and variability of substratum movement (adapted from the definition of flow...
stability by Poff and Ward, 1989). A milder physical disturbance regime implies less scouring of the substratum and less didymo biomass removal. As such, I predicted low didymo standing crop and cover in unstable reaches and high didymo standing crop and cover in highly stable reaches. I found support for this prediction in that streambed stability – measured using the Pfankuch index – explained 78% of the variation in didymo standing crop and 70% of the variation in didymo cover. Both effect sizes are for didymo-positive sites \( n = 9 \). These results suggest that Pfankuch score is an important habitat variable controlling didymo mat formation.

Few other researchers have examined the relationship between bed stability and didymo biomass. Kilroy and Bothwell (2012) found a marginally significant relationship between the proportion of substratum less than 6cm in diameter and didymo standing crop across 15 reaches \( r^2 = 0.27 \). This relatively low \( r^2 \) value may be attributable to the fact that substratum size is only one element of bed stability at the reach scale. In contrast, the Pfankuch score includes the bed stability elements of substratum size, substratum shape, particle consolidation and patterns of scour and deposition (Pfankuch, 1975).

It is worth considering that stream flora are capable of altering streambed stability under certain conditions. For example, Fritz et al. (2004) performed a manipulative experiment in a third-order stream in Alabama. They measured substratum stability several months after removing stems and roots of the herbaceous aquatic plant Justicia americana. They found that the force needed to shift stones was higher where J. americana was left intact. However, it is difficult to apply this result to didymo because J. americana has a deeply-rooted growth form but didymo attaches by stalks to the surfaces of stones.

If didymo biomass does alter streambed stability, the effect is likely to be vanishingly small. At the low end of the streambed stability spectrum, didymo can not increase bed stability because the substratum moves rarely, if ever. At the other end of the spectrum, sites with harsh physical disturbance regimes generally prevent the growth of sufficient didymo biomass for positive feedback effects to be measured. For rivers with moderate physical disturbance regimes (ie. between the two extremes), positive feedback can only occur when the presence of didymo biomass results in less substratum movement. If didymo grows by filling the spaces between cobbles, for example, a larger flood may be required to move the bed. However, didymo biomass initially accumulates on the surfaces of stones, rather than between them (Whitton et al., 2009). Also, rather than lowering drag, didymo-covered rocks were shown to experience higher drag compared to bare cobbles (Larned et al., 2011). Therefore, positive feedback effects between streambed stability and didymo biomass are unlikely in New Zealand rivers.
Comparison of predictors

The results of this study give moderate support to my hypothesis that Pfankuch value is a better predictor of didymo standing crop and cover than dissolved reactive phosphorus concentration or turbidity. For sites with established didymo mats, models including Pfankuch value were clearly separated from all other models by AICc. Models fitted to the full dataset, however, were unable to be clearly separated because the $\Delta$ AICc values were small (less than seven).

In the results, the effect of dissolved reactive phosphorus on didymo standing crop was weaker than in previous studies. Across a survey of 15 reaches, for example, mean dissolved reactive phosphorus concentration explained 49% of the variation in didymo standing crop (95% CI from 0.06 to 0.82; Kilroy and Bothwell, 2012). The authors of this study also showed that, at the confluence of two rivers with different phosphorus concentrations, didymo mats grew to higher biomass on the side with lower phosphorus. The 95% confidence intervals for the proportion of variation in didymo standing crop explained by phosphorus concentration in the survey conducted by Kilroy and Bothwell (2012) and in my survey were relatively wide. This indicates that the power of both studies was low and could be improved by increasing the number of sites surveyed.

The above results need to be interpreted with caution due to two limitations of the study. Firstly, the variability in phosphorus and turbidity at each site remains unknown because measurements were taken only once. Kilroy and Bothwell (2012) used measurements taken over several years from a monitoring programme instead. However, the phosphorus results from my study and Kilroy and Bothwell (2012) were similarly distributed (Figure 2.5).

![Figure 2.5: Comparison between this study and Kilroy and Bothwell (2012, KB) for dissolved reactive phosphorus (A) and log-transformed standing crop (B). Measurements of the same value are stacked horizontally.](image-url)
In both studies, the distribution of phosphorus concentrations were centred around two parts per billion and greater than 90% of measurements were below eight parts per billion. Despite this similarity, taking several measurements would capture some of the within-site variability in dissolved reactive phosphorus and turbidity allowing more powerful regression techniques, such as general linear mixed models, to be used in the analysis.

Secondly, measures of didymo biomass were taken at a single point in time rather than at the peak of biomass accumulation. This is likely to result in an underestimate of the biomass that can be supported at a particular site. Future studies of the link between site characteristics and didymo bloom formation would benefit from using measurements of the peak biomass attained at each survey location.

Despite these limitations, Pfankuch index appears to be a strong predictor of didymo standing crop and cover at the reach scale. These results indicate that physical disturbance regime is a primary habitat variables controlling the biomass of didymo. This conclusion is useful for managers to aide in the identification of reaches susceptible to didymo blooms.
Chapter 3

Modelling the distribution of *Didymosphenia geminata* blooms in New Zealand

3.1 Introduction

In nature, the distribution of a species is determined by countless factors which have previously been characterised as follows (Soberon and Peterson, 2005; Jimenez-Valverde *et al.*, 2011):

1. abiotic factors - the physical environment of a species including nutrient availability, disturbance regime and light intensity
2. biotic factors - interactions between species such as competitive exclusion
3. dispersal limitation - includes the presence of barriers to dispersal and the ability of a species to disperse
4. evolutionary factors - capacity to adapt to new conditions.

The fundamental niche, potential niche and realised niche of a species each incorporate one or more of the above categories to form a hierarchy of niche concepts. The fundamental niche is described by the full set of abiotic factors constraining a species. Because there are usually many abiotic factors, the fundamental niche is conceptualised as a multidimensional space within which the values of each abiotic factor permit growth and reproduction (Hutchinson, 1957). The multidimensional fundamental niche can be projected onto a map of a study region if information on each abiotic
factor is available there (Jimenez-Valverde et al., 2011). Such a map describes the potential niche for a species. Whereas the fundamental niche contains all the combinations of abiotic factors allowing growth and reproduction, the potential niche does not. That is because some conditions in the fundamental niche are never found in the environment (Jackson and Overpeck, 2000). For example, the toxin-producing alga *Prymnesium parvum* can grow at 20°C in an artificial lake water medium with 3g/L of NaCl (Baker et al., 2009) but these conditions are not found in a hypothetical study area (a cold, freshwater lake for example). In this case, as in all cases, the potential niche of the species is smaller than its fundamental niche.

There is a third niche concept to consider – the ‘realised niche’. The realised niche is a subset of the potential niche within which all biotic factors relevant to a species are satisfied (Soberon and Peterson, 2005). These biotic factors include the presence of beneficial interactions (e.g. biofilms for attachment) and the absence of all strong competitors, specialised predators and diseases. Even though the realised niche takes into account all abiotic and biotic factors, it may not perfectly describe the actual distribution of a species. That is because the actual distribution of a species is constrained by dispersal limitations and influenced by the capacity of the species to adapt to new conditions as well (Soberon and Peterson, 2005).

**Species Distribution Modelling**

In species distribution modelling, species occurrence data and a set of environmental layers are used to produce models of the potential niche, realised niche or actual distribution of a species (Elith and Leathwick, 2009). It is possible to tell which niche concept is the focus of a species distribution model by the environmental layers included within it. Abiotic environmental layers indicate that the potential niche is being modelled; a combination of abiotic and biotic environmental layers indicate that the realised niche is being modelled; and environmental layers representing all four categories (abiotic, biotic, dispersal limitations and evolutionary) indicate that the actual distribution is being modelled.

Species distribution modelling has been increasingly used to estimate the potential niche of invasive species (Mau-Crimmins et al., 2006; Peterson and Nakazawa, 2008; Jimenez-Valverde et al., 2011).

There are two main approaches to species distribution modelling: mechanistic and correlative. The mechanistic approach consists of measuring the response of individuals to a set of abiotic variables to identify the range of values permitting growth and reproduction (Guisan and Zimmermann, 2000). Suitable regions in geographic space can
then be identified. The second approach – correlative species distribution modelling – consists of relating occurrence data to environmental layers with a particular algorithm (such as logistic regression). A diverse array of algorithms exist for this purpose, three of which I will outline and use in this chapter.

A considerable challenge for the correlative approach is to make a reliable estimate of the potential niche of a species given occurrence data from its actual distribution. As already outlined, the actual distribution is generally not coincident with the potential niche because the potential niche does not include dispersal limitations or the capacity of the species to adapt to new conditions. This challenge can be met – at least in part – by examining model response functions to check that they are similar to what may be expected for the species (Elith et al., 2010) and testing the fitted models on withheld data (Elith and Leathwick, 2009).

Species Distribution Modelling Algorithms

There are many correlative species distribution modelling algorithms available to ecologists. Logistic regression has been widely used for several decades (Nash and Bradford, 2001). Splitting ‘feature space’ with boosted regression trees, and machine learning with Maxent have performed well in comparative studies (Elith et al., 2006; Phillips et al., 2009). Here I provide a brief description of these three methods.

Logistic regression is a widely-used technique in species distribution modelling, in part because it is familiar to many biologists. The logistic model relates a linear sum of predictors \( \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \) to the probability of success \( \pi \) via the logit transformation \( \ln(\frac{\pi}{1-\pi}) \) where \( \ln \) is the natural logarithm. In species distribution modelling, the predictors are the values of the environmental layers at the sample locations (presences and absences) and the probability of success is interpreted as the probability of presence.

Boosted regression trees rely principally on the successive splitting of ‘feature space’ – the possible combinations of values for the environmental layers. Each split adds a branch to a regression tree. Multiple regression trees are combined in a step-wise process called stochastic gradient boosting that has similarities with model averaging (Elith et al., 2008). Stochastic gradient boosting improves model accuracy and reduces over-fitting (where the noise in the data is modelled instead of the overall patterns or signal).

Maxent is a software package that applies the maximum entropy principle to species distribution modelling (Phillips et al., 2006). Maxent takes, as input, a list of presence locations for a species – it does not require absences – and a set of environmental layers.
Table 3.1: An example of model sensitivity and specificity with two presences (filled circles) and two absences (hollow circles). Sensitivity is the proportion of presences in the actual state that are presences in the model prediction. Specificity is the proportion of absences in the actual state that are absences in the model prediction.

<table>
<thead>
<tr>
<th>Actual State</th>
<th>Model Prediction</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>●●●●</td>
<td>●●●●</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>●●●●</td>
<td>●●●●</td>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>●●●●</td>
<td>●●●●</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>●●●●</td>
<td>●●●●</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The environmental layers carry values for environmental variables at a set of points (or river sections in this case). Maxent produces a value for all points such that the distribution of those values has maximum entropy given a set of constraints derived from the environmental layers. The set of values assigned to the points may, under restrictive conditions, be interpreted as the relative probability of species presence.

To date, no study of the distribution of didymo in New Zealand has compared the predictions produced by different species distribution modelling algorithms. The principle advantage of using multiple algorithms is the ability to examine the influence of algorithm choice on the resulting prediction maps.

Distribution Modelling of Didymo

Didymo is a benthic stalk-forming diatom introduced to the South Island of New Zealand during or before 2004 (Kilroy and Unwin, 2011). It has spread to several hundred waterways and is capable of forming thick benthic mats (refer General Introduction and Chapter 2 for more details of didymo’s ecology and rapid spread).

A robust distribution model for didymo is an important tool for managing its spread. In the US, Kumar et al. (2009) used 26 environmental layers representing climate, topography, geology and hydrology to model the potential distribution of didymo. They compared the prediction maps from four different distribution modelling algorithms: Logistic regression, maximum entropy modelling (Maxent; Phillips et al., 2006), classification and regression trees (CART; De’ath and Fabricius, 2000) and generalised additive models (GARP; Stockwell, 1999). In terms of model performance, the Maxent model had the highest sensitivity and specificity (refer Table 3.1 for definitions of these model performance measures).

In New Zealand there have been two efforts to model the potential distribution of
didymo. In the first of these, Kilroy et al. (2008) chose six abiotic variables which were likely to be important for didymo growth. The variables were the proportion of the upstream catchment with glaciers (a light-availability surrogate because glaciers tend to contribute to the turbidity of streams by acting as a source of suspended sediment), the proportion of the upstream catchment with peat (peat alters the water chemistry of streams by contributing dissolved organic carbon and hydrogen ions), the proportion of the upstream catchment with lakes (lakes typically reduce stream flow variability), days per month with rain greater than 25mm (another flow variability surrogate), the hardness of catchment rocks (harder rocks may remain immobile during high flows, reducing disturbance removal of algal biomass) and summer air temperature (temperature is an important variable for periphyton). For each variable, the authors chose an optimum value for didymo growth based on scant physiological data for the species. They then calculated the Gower metric (an index of similarity) between the optimum values and the conditions found in each New Zealand river segment, interpreting higher values for the Gower metric as being less suitable for didymo growth. A weakness of their approach is that the Gower metric is sensitive to the choice of optimum values. For instance, their own sensitivity analysis showed that a modest change in the chosen optimum summer air temperature by 3°C increased the proportion of river segments in the high-risk category 24-fold (0.4% to 9.7%). A more robust approach – while physiological data remains scarce – is to build a distribution model based on presence and absence locations (the river segments where didymo has been recorded and has never been recorded, respectively). There were, however, too few presence locations at the time this study was conducted to use this approach.

In the second effort to model the distribution of didymo in New Zealand, presence and absence locations were considered. Rather than constructing a model directly from presences and absences though, Kilroy et al. (2007) based their model on a survey of didymo cover and didymo mat thickness in 17 rivers at the reach scale (40 to 50m). The authors combined this survey data with 18 environmental layers available across New Zealand. The modelling algorithm used was boosted regression trees (BRTs Elith et al., 2008). They found that Lake buffering upstream (a flow stability variable) and weighted bed sediment classes (a streambed stability variable) were the two most important predictors of didymo mat thickness. As numerous records of visible didymo mats have been collected (through the Didymo Samples Database MAF Biosecurity NZ, 2012), there is an opportunity to extend this work by including many more rivers in the modelling process. By including more sites, a clearer picture of where didymo is able to form blooms in New Zealand rivers will emerge.
Pfankuch Bed Stability and Didymo

Streambed stability is an important abiotic factor for periphyton (Biggs, 1995) but the potential for this factor to be a useful predictor of didymo presence has not been fully explored. The Pfankuch bed stability index is a method for evaluating streambed stability from a visual-assessment of a reach. It is much less laborious to perform than previous measures of bed stability but produces reliable estimates (Schwendel et al., 2011). In Chapter 2, I found Pfankuch bed stability to be an important predictor of didymo standing crop and percentage cover. Unfortunately, a Pfankuch bed stability variable covering all river segments in New Zealand is not presently available. There might be, however, a suitable surrogate variable for Pfankuch bed stability within an existing spatial database covering all New Zealand river segments.

Chapter Aims

My aims in this chapter are:

1. To examine the correlations between Pfankuch index values from my bed stability survey (Chapter 2) and a suite of environmental variables. These variables must be available as environmental layers covering all New Zealand rivers.

2. To model the potential distribution of didymo in New Zealand using a set of environmental layers directly related to the growth of didymo. I will carry out distribution modelling using three distribution modelling algorithms: Logistic regression, boosted regression trees and Maxent.

3. To create and compare geographic projections of the three models developed in Aim 2 for the North and South Islands of New Zealand. Comparisons will be made in terms of model performance, model predictions (maps) and model structure.
3.2 Methods

Environmental Layers

Environmental layers for modelling the distribution of didymo were sourced from two spatial geodatabases: Freshwater Ecosystems of New Zealand (FENZ) and WorldClim. The FENZ database covers all New Zealand rivers, lakes and wetlands. The dataset for rivers within FENZ contains 34 environmental layers (four of which are transformations of other variables). FENZ was developed by the Department of Conservation with input from the National Institute of Water and Atmospheric Research and Landcare Research (Leathwick et al., 2010).

WorldClim is a global climate database (Hijmans et al., 2005). It contains climate layers at a resolution of 30 arc-seconds interpolated from weather station data. The database also contains 19 bioclimactic variables derived from monthly minimum temperature, maximum temperature, precipitation and altitude variables.

I excluded layers in the FENZ database which were transformations of other variables. For example, I kept the layer representing mean annual 7-day low flow but removed the fourth root transformed version. Then, I chose nine environmental layers from the FENZ and WorldClim databases to use for modelling. The nine variables each have a plausible, direct link to the growth or colonisation of didymo. Using direct (proximal) environmental variables, rather than indirect (distal) environmental variables leads to the construction of more ecologically relevant models (Elith and Leathwick, 2009). Several variables were included because they were important predictors in previous didymo distribution modelling efforts. Lake buffering upstream (USLake) and weighted bed sediment classes (ReachSed) were the top two most important predictors in the didymo mat thickness model developed by Kilroy et al. (2007). Other variables were chosen because they were similar to important variables in other studies. Mean air temperature in the warmest month (SegJanAirT) and days of rainfall greater than 25mm (USDaysRain) were chosen for this reason; being similar to the best predictors of didymo occurrence in Kumar et al. (2009). Descriptions of the environmental layers are given in Table 3.2.

I created an additional variable, called DAYS_BTW_F, representing the mean number of days between floods three times median flow. Flows of this magnitude are considered to be important for benthic communities through physical disturbance of the streambed (Clausen and Biggs, 1997). The calculation for DAYS_BTW_F follows Kilroy et al. (2007). Briefly, I divided the number of days in a year by the annual frequency of floods. Flood frequency was estimated for each river segment in the FENZ
network using two categorical variables (SRC_OF_FLW and CLIMATE). To convert the categories to flood frequencies, a conversion chart from the river classification study by Snelder et al. (2005) was used.

**Pfankuch Index Values**

The Pfankuch index of riverbed stability was measured at forty sites in the South Island of New Zealand for a survey of didymo standing crop and cover (Chapter 2).

Five environmental layers were chosen as potential surrogates for Pfankuch bed stability. The proportion of riparian shading (SegRipShad) was chosen because well-vegetated upper banks promote bed stability. Bed sediment classes weighted by proportional cover (ReachSed) was included because streambeds with coarser sediments are more stable (Matthaei et al., 1999a). The remaining three potential surrogates for Pfankuch bed stability – lake buffering in the upstream catchment (USLake), days per year with rain greater than 25mm (USDaysRain) and mean number of days between floods 3x median flow (DAYS_BTW_F) – are variables that characterise the flow regime. This fact is relevant because streambeds tend to be more stable when the flow regime is more stable (Death and Winterbourn, 1994).

To get values for the potential surrogate variables at the same locations where Pfankuch bed stability was measured (at the sites surveyed in Chapter 2), the nearest river segment in the FENZ network was selected. The values of the environmental layers at all forty survey sites (the sites surveyed in Chapter 2) were extracted using ArcMap (ESRI, 2012). Then, Pearson and Spearman correlation coefficients were calculated between the extracted values and Pfankuch index values.

**Didymo Occurrence Data**

Occurrence data of Didymo in New Zealand were sourced from the Didymo Samples Database maintained by MAF Biosecurity New Zealand MAF Biosecurity NZ (2012). The Didymo Samples Database contains didymo presence/absence records from surveys conducted by regional councils, the National Institute of Water and Atmospheric research, and other agencies in New Zealand. All records between January 2005 and July 2013 (5569 in total) were downloaded on 31 July 2013 and sorted by visit date. Each record contained information on the presence or absence of didymo colonies or cells in benthic and drift samples (see Table 3.3).
Table 3.2: Environmental layers used to model the potential distribution of didymo. The names of each layer are in bold with the data source in brackets below. The data sources are The Freshwater Ecosystems of New Zealand database (FENZ) or the global climate database (WorldClim). The exception is DAYS_BTW_F, which was calculated from the river environment classification in FENZ and flood frequency estimates from Snelder et al. (2005). Where transformations were applied, they are shown in brackets. The expected response is a brief description of the expected change in didymo occurrence probability with increasing values of the environmental variable.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Variable and Expected Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegRipShad</td>
<td>Estimate of riparian shading (proportion)</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>USGlacier</td>
<td>Glacial cover in upstream catchment (proportion)</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>SegCluesN</td>
<td>Nitrogen concentration in ppb (log x+1 transformed)</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>SegJanAirT</td>
<td>Mean January air temperature</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>BIO4</td>
<td>Temperature seasonality</td>
</tr>
<tr>
<td>(WorldClim)</td>
<td></td>
</tr>
<tr>
<td>ReachSed</td>
<td>Bed sediment classes weighted by proportional cover</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>USLake</td>
<td>Lake buffering in upstream catchment</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>USDaysRain</td>
<td>Days per year with rain &gt;25mm in upstream catchment</td>
</tr>
<tr>
<td>(FENZ)</td>
<td></td>
</tr>
<tr>
<td>DAYS_BTW_F</td>
<td>Mean days between 3x median flow (log-transformed)</td>
</tr>
<tr>
<td>(Calculated)</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3: Variables in the Didymo Samples Database indicating presence or absence of didymo colonies and cells. The first four variables are for different sampling methods while the final variable is derived from combining the findings from ‘Benthic Result’ and ‘Drift Result’.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Interpretation</th>
</tr>
</thead>
</table>
| Didymo Visible | Were didymo colonies visible on the benthos?  
Levels: Yes, No, Unknown, Suspect |
| Benthic Result | Presence of didymo cells in benthic samples  
Levels: Positive, Negative, Unknown |
| Drift Result   | Presence of didymo cells in plankton net sample  
Levels: Positive, Negative, Unknown |
| Visual Result  | Presence of didymo cells in additional periphyton samples  
Levels: Positive, Negative, Unknown |
| Visit Result   | Positive: Benthic Result OR Drift Result were positive  
Negative: Benthic Result AND Drift Result were negative |

Cleaning of Didymo occurrence records was necessary to remove spurious positives and purge records with missing information. Cleaning proceeded as follows:

1. Reclassified seven records with a Benthic Result, Drift Result or Visual Result from ‘flag for recheck’ to ‘negative’ or ‘positive’ based on information provided by Phoebe Zarnetske (NIWA, personal communication). I removed one record which could not be classified.

2. Changed 11 records on the North Island spuriously listed as ‘positive’ for Benthic Result, Drift Result or Visit Result to ‘negative’. This was done because, to date, no report of didymo has been confirmed on the North Island.

3. Removed two records with Didymo Visible recorded as ‘positive’ but Benthic Result ‘negative’. This conflict suggests that didymo colonies were recorded by mistake.

4. Further, I removed 24 records which were missing valid site coordinates and 48 records which were missing a visit date.

Following data cleaning, occurrence records were classified into presence and absence groups based on the columns in Table 3.3. I aimed to model the potential distribution of didymo colonies. It follows that presence was defined as the presence
of didymo colonies, rather than cells. Absence was defined as the absence of didymo colonies AND didymo cells. Checking for the absence of both colonies and cells helped to eliminate false-negatives.

During the classification step, I considered the sources of false-positives. One type of false-positive occurred when algal colonies were misidentified. I detected these false-positives by comparing the Didymo Visible, Visual Result and Benthic Result columns of the Didymo Samples Database. When Didymo Visible was ‘Yes’ but Benthic Result or Visual Result were ‘Negative’, the Didymo Visible result was treated as a false-positive and the record then classified as an absence. A second type of false-positive occurred when cells were transported from upstream colonies and settled on the bed (Kilroy and Dale, 2006). To minimise these false-positives, I used the Didymo Visible column to verify the Visual Result and Benthic Result columns. When Visual Result or Benthic Result were ‘Positive’ and Didymo visible was ‘Unknown’ or ‘No’ I chose not to classify that record due to this conflict. The steps I used to classify records are shown in Figure 3.1.

```
One record
|
| Visit Result was ‘Negative’............................... Absence
|
| Otherwise, continue
|
| Didymo Visible was ‘Yes’ OR ‘Suspect’
| |
| Benthic Result was ‘Negative’....................... Absence
| |
| Benthic Result was ‘Positive’............... Presence
| |
| Benthic Result unknown
| |
| Visual Result was ‘Negative’................. Absence
| |
| Visual Result was ‘Positive’............. Presence
| |
| Visual Result unknown................ Presence
|
| Didymo Visible unknown OR ‘No’
| |
| Benthic Result was ‘Negative’............... Absence
| |
| Benthic Result was ‘Positive’........ Discard
| |
| Benthic Result unknown
| |
| Visual Result was ‘Negative’........ Absence
| |
| Visual Result was ‘Positive’........ Discard
| |
| Visual Result unknown........ Discard
```

Figure 3.1: Classification scheme applied to didymo occurrence records in the Didymo Samples Database.
Because didymo is a recently-introduced species to New Zealand and there are barriers to its spread – mountains between catchments and perhaps the Cook Strait – it is likely that it is not in distributional equilibrium (in all the places it could be). To reduce the influence of the non-equilibrium state of didymo in New Zealand on the model outcomes I excluded North Island absence records from the analysis.

Mapping Records to River Segments

A map of the New Zealand river network was created in ArcMap (ESRI, 2012). First, river segments of stream order three or greater were selected from the Freshwater Ecosystems of New Zealand database (Leathwick et al., 2010). Didymo has not been recorded in reaches with stream order less than three (Kilroy et al., 2007).

To standardise the coordinate system, coordinates from the occurrence data were converted from NZTM2000 to WGS1984 using an online conversion tool (Land Information New Zealand, 2013).

The records with standardised coordinates were plotted on the New Zealand river network map. On inspection, some records fell slightly outside a pixel on the river map either due to inaccuracies in the recorded GPS locations or inaccuracies in the river network map. Records falling more than 15 arc-seconds (approximately 500m) from a reach were deleted. Records falling within 15-arc-seconds of a reach were associated with the nearest reach. The cut-off distance (15 arc-seconds) is somewhat arbitrary - half the map resolution in this case. Distribution modelling algorithms are generally robust to changes in the occurrence locations of this magnitude (Graham et al., 2008).

Multiple records fell on many of the river segments. To classify the pixels as ‘presence locations’ or ‘absence locations’ I applied two rules. If ALL the records on a given segment were classified as Absences, the segment was treated as an absence location. If ANY record on a given segment was classified as Present, the segment was treated as a presence location. The segments were classified using R with accuracy checked by detailed examination of a randomly chosen set of 50 segments.

I identified reliable absences by two methods. (1) Selected absence locations downstream of presence locations. I reasoned that, due to the tendency for didymo cells to be transported long distances downstream but only short distances upstream, absences downstream of presence locations are less prone to false-negatives. To further reduce false-negatives, I checked the visit date for downstream sites to see if the last visit was after the first-recorded upstream presence. If the visit date was after the first upstream presence, I classed that location as a reliable absence. (2) Selected absence locations with a positive drift net result. I considered these absences to be reliable because a
positive drift net result indicates that cells were detected in the water column. In that
case, I reasoned that didymo cells have been introduced to that location but did not
establish. Both types of absences - those downstream of previously-recorded presences
and those with a positive drift net result - were pooled together.

To reduce spatial autocorrelation (the tendency for points which are close in geo-
graphic space to have similar values for an environmental variable) in the presence and
absences locations, I plotted the survey locations on a map of the New Zealand river
network. In places where the didymo survey sites were clustered over short distances
(less than 10km) on the same rivers, I removed site clusters by randomly retaining one
survey location from each set.

Training and Testing Datasets
The didymo survey data were partitioned into two groups so that cross-validation could
be performed. Cross-validation is a method for testing the performance of predictive
models. It involves splitting the original dataset into training and testing groups.
The training group is fed into the model to calculate the various weightings, features
or coefficients. The testing group is used to check the model performance by letting
the model make predictions for each site in the group and then calculate a set of
performance measures (Guisan and Zimmermann, 2000). I used half of the didymo
presence locations for training and half for testing, the same ratio of training to testing
presences used by Kumar et al. (2009). This ratio allowed a reasonably large sample
size for model training (51 presence records).

Logistic Regression Model Fitting
All models were fitted using R (R Core Team, 2013). I began by fitting a global model
with the nine environmental variables (glm, binomial family). I then ran an automated
model selection function (dredge from the mumin package by Barton, 2013) allowing
up to seven predictors with model ranking by AICc (refer Chapter 2 for details of
AICc). Models within the top 2 AICc values (the top models) were isolated from
the full set of fitted models. A threshold of 2 was used because models with ∆AICc
values in this range are difficult to separate in terms of fit and complexity (Burnham
et al., 2011). A higher threshold of 7, was first applied but was rejected because a
high proportion of the fitted models were included. The ‘nesting rule’ (Richards et al.,
2011) was applied to the top models. That is, any model with more parameters than
a higher-ranked model was removed. Model averaging was unnecessary because, after
applying the nesting rule, only a single top model remained.
The contribution of variables to the final model was calculated from the drop in deviance attributable to each variable. That is, the drop in deviance between the null model (no predictors) and models containing an isolated variable from the final model (one predictor). Variable contribution was reported as the proportional drop in total deviance.

The proportion of deviance explained ($D^2$) was calculated by dividing the drop in deviance by the null deviance. The null deviance is calculated for the model with no predictors - giving a constant fitted value equal to the mean of the response. The proportion of deviance explained is a measure of effect size. A value for $D^2$ of 0 indicates that the model is no better than the null model while a value of 1 indicates the model fits the training data perfectly (Leathwick et al., 2006).

$$D^2 = \frac{\text{nullDeviance} - \text{residualDeviance}}{\text{nullDeviance}}$$

**BRT Model Fitting**

I fitted boosted regression trees following Elith et al. (2008) using `gbm.step` in the dismo R package (Hijmans et al., 2013). The modelling variables I set prior to running the algorithm were ‘tree.complexity’, ‘learning.rate’ and ‘bag.fraction’. Tree complexity is an integer expressing the ‘interaction order’ of the data (Elith et al., 2008). When tree.complexity equals 1, no interactions are fitted; when tree.complexity equals 2, first-order interactions are fitted; when tree.complexity equals 3, second-order interactions are fitted and so on. The value chosen for tree.complexity should reflect the sample size and expected interaction order of the training dataset. As a rough guide, at least 250 sites are needed to reliably fit second-order interactions (Elith et al., 2008). Learning rate is a multiplier applied to each tree which shrinks the contribution it makes to the final model predictions. Smaller learning rates lead to a higher number of trees being fitted and improved model performance (Hastie et al., 2009b). Typically, learning.rate is smaller than 0.01. Bag fraction is the proportion of training data used to fit each tree in the model. This proportion is drawn at random in each iteration and helps to improve the predictive performance of the model. Typical values for bag.fraction are between 0.5 and 0.75 (Elith et al., 2008).

I was interested in first-order interactions so set tree.complexity to 2. Higher values were not tried due insufficient sample size (less than 250 sites in the training dataset; Elith et al., 2008). The optimal number of trees was determined using 10-fold cross-validation (Hastie et al., 2009a). I adjusted the learning rate to give at least 1000 regression trees, settling on a value of 0.002. I tried bag.fractions of 0.5,
0.6 and 0.7 separately then examined the drop in deviance. In the final model, I set bag.fraction to 0.5. I attempted to simplify the model by dropping up to 5 predictors using `gbm.simplify` (Hijmans et al., 2013). Single variable response plots were produced using the `gbm.plot` function.

**Maxent Model Fitting**

Maximum entropy models were fitted using the java implementation of Maxent (Dudik et al., 2011) and the dismo R package (Hijmans et al., 2013). I started by fitting a global model with all nine predictors to the training presences, specifying all training absence locations as background samples. This approach to background sampling is discussed further in Dudik et al. (2006). I specified that the algorithm fit linear, quadratic and hinge feature types - as recommended by Phillips and Dudík (2008) given the sample size in my study. I used the default regularisation parameters.

I then examined the drop in gain when each variable was excluded in turn from the global model (using the jackknife option in Maxent). A larger drop in gain indicated that the excluded variable had more information not present in the other variables (Dudik et al., 2011). A simplified model was constructed using the variables associated with a drop in gain greater than 0.01. The simplified model was fitted to the same training presences as the global model using the same background locations, feature types and regularisation parameters. The relative contribution of variables to the simplified model was recorded from the Maxent output. Single variable response plots were produced using the `response` function.

**Performance Measures**

Model performance was evaluated using three performance measures: sensitivity, specificity and the area under the receiver operating characteristic curve (AUC). Sensitivity, in species distribution modelling, is the proportion of known presences that a model classifies as ‘present’ in the testing dataset. Specificity is the proportion of known absences that a model correctly classifies as ‘absent’ in the testing dataset (for clarification of the difference between these two performance measures refer Table 3.1). Sensitivity and specificity are threshold-dependent measures because they require the user to specify a classification level (Fielding and Bell, 1997). A classification level (or threshold) is the point where the model predictions (ranging from 0 to 1) are grouped into presences and absences (0 or 1 only).

I used two classification thresholds for all the sensitivity and specificity calculations. The first threshold was chosen to maximise the sum of sensitivity and specificity. This
threshold was found by repeatedly calculating sensitivity and specificity for all thresholds between 0 and 1 with an increment of 0.01 and then finding the highest value of sensitivity plus specificity. I used this particular classification threshold because it has been shown to be relatively insensitive to prevalence (the proportion of presences) in the training data Liu et al. (2005).

The second threshold was chosen such that specificity equals approximately 0.5. I used this classification threshold because it permits the sensitivity of the models to be compared under equal specificity. In effect, I am asking ‘if I allow the model to misclassify 50% of absences, how well does it classify presences?’ This is a useful question because it is usually better for a species distribution model to make reliable predictions of an unwanted organism’s presence rather than its absence. That is because the consequences of its presence are potentially great (in terms of changes to the ecological community) while its absence has no ecological impact (it was never there, it is still not there).

In contrast to sensitivity and specificity, AUC (or Area Under Curve) is a threshold-independent measure of performance. AUC was calculated by plotting 1-specificity and sensitivity across the range 0-1 and then determining the area under the curve using `gbm.roc.area` in the gbm R package (Ridgeway, 2013). The range of values for AUC are between 0.5 (which indicates performance no better than random) and 1 (which indicates perfect prediction).

**Prediction Maps**

Prediction maps were created in ArcMap (ESRI, 2012). The predictions were linked to individual river segments in the FENZ network by their NZREACH number (a unique number assigned to each reach). River segments were then colour-coded using the following scheme: Predictions between zero and the classification threshold where specificity approximately equals 0.5 (Threshold 2) were coloured dark green. Then, the highest prediction value was found (pMax) and the difference between pMax and the Threshold 2 was broken into three equal intervals. River segments were coloured light green if their predicted value fell in the first interval, orange if in the second interval and red if in the third interval. Using this colour scheme, dark green river segments have a relatively low risk of didymo bloom formation and red river segments have a relatively high risk of didymo bloom formation. To improve clarity, region boundaries from the New Zealand Digital Boundaries Dataset (Statistics New Zealand, 2008) and a coastline shapefile from Land Information New Zealand (LINZ, 2013) were added.
3.3 Results

Pfankuch index correlations

To address the first aim of this chapter, I calculated Pearson’s and Spearman’s correlation coefficients for Pfankuch index value and nine environmental layers identified as being relevant to the ecology of didymo from the Freshwater Ecosystems of New Zealand (FENZ) database. As Table 3.4 shows, the layers with highest Spearman correlation coefficients were lake buffering in the upstream catchment (USLake) and days per year with rain greater than 25mm in the upstream catchment (USDaysRain). These results should be interpreted with caution because, when USLake was zero (32 of 40 data points), Pfankuch values covered a wide range of values (see Figure 3.2). Although USDaysRain had no zero values, the points are widely-scattered around the regression line.

Table 3.4: Pearson’s and Spearman’s correlation coefficients for Pfankuch bed stability values and environmental layers included in this study. Values for the environmental layers were taken from the nearest river segment to the survey location where Pfankuch bed stability was measured (N=40).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pearson’s r</th>
<th>Spearman’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegRipShad</td>
<td>0.110</td>
<td>0.089</td>
</tr>
<tr>
<td>ReachSed</td>
<td>0.056</td>
<td>0.036</td>
</tr>
<tr>
<td>USLake</td>
<td>-0.613</td>
<td>-0.489</td>
</tr>
<tr>
<td>USDaysRain</td>
<td>-0.286</td>
<td>-0.304</td>
</tr>
<tr>
<td>DAYS_BTW_F_T</td>
<td>-0.476</td>
<td>-0.135</td>
</tr>
</tbody>
</table>

Distribution Modelling Datasets

There were 5569 records in the Didymo Samples Database as at 31 July 2013. Approximately 10% of these records were removed during data cleaning, leaving 5010 records. Applying the classification scheme described in Figure 3.1 yielded 4869 absence records and 141 presence records (97% and 3%, respectively). When mapped to the nearest river segment, 4702 records were within 0.00417 decimal degrees, leaving 308 records too far away (just over 6%). These 4702 mapped records fell on 1100 unique locations. Applying my classification rules to the unique locations resulted in 975 absence locations and 125 presence locations. After removing site clusters and excluding absence
locations on the North Island, 517 absence locations and 101 presence locations remained. I identified 39 of the absence locations as ‘Reliable Absences’. The reliable absences were either downstream of previously-recorded presences or had a positive drift-net result.

Two datasets were used to fit (or ‘train’) the distribution models. Training 1 contained 39 reliable absences and 51 presences while Training 2 contained 200 absences and the same set of 51 presences (locations shown in Figure 5). In distribution modelling it is common to compare the range of values covered by the environmental variables at the training locations. From the violin plots (Appendix C), it was apparent that Training 1 and Training 2 covered a similar range of values for six of the environmental layers but not for the remaining three. Training 2 covered higher values for the proportion of glacial cover in the upstream catchment (USGlacier) and nitrogen concentration (SegCluesN) and lower values for bed sediment classes weighted by proportional cover (ReachSed). It is also apparent that the two training datasets covered the full range of values found at all didymo survey locations (‘All NZ’ in the figures) remarkably well. Notable exceptions are USGlacier, SegCluesN and lake buffering in the upstream catchment (USLake) for which the training datasets covered the lower values found in New Zealand but missed the high end of the range.
Distribution Model Fitting

An early decision I needed to make in the distribution model fitting stage was whether to use the Training 1 or Training 2 datasets. I assumed that there would be a benefit in using the reliable absences in Training 1 (that is, absences downstream of presence locations or with a positive drift net result) over the full list of absences in Training 2. To find out, I fitted a logistic regression model to Training 1 and calculated the proportion of deviance explained ($D^2$). I then repeatedly drew 39 absences from Training 2 and again calculated ($D^2$). Using this bootstrapping approach, I saw that ($D^2$) for a model fitted to Training 1 was lower than the mean ($D^2$) value from the set of models fitted to Training 2. This result showed that there was no apparent benefit in using the ‘reliable’ absences in Training 1.

Using the Training 2 dataset, I addressed the second aim of this chapter by fitting distribution models using three distribution modelling algorithms. They were logistic regression, boosted regression trees and Maxent. For all three algorithms, I fitted a global model (containing all environmental layers) and a simplified model using appropriate model selection methods for each algorithm. I refer to the simplified models as the ‘final models’ because they were used to produce the didymo distribution prediction maps.

The final logistic regression model retained three variables. They were temperature seasonality (BIO4), nitrogen concentration (SegCluesN) and proportion of upstream catchment with glaciers (USGlacier). For river segment $x$, the predicted value $p_x$ according to the logistic regression model was:

$$
\text{logit}(p_x) = -8.96 + 0.00221\text{BIO4}_x - 1.89\text{SegCluesN}_x - 0.138\text{USGlacier}_x
$$

The signs of the coefficients indicate the direction of relationship (but their magnitudes can not be directly compared because the data were not standardised prior to model training). In this model equation, the probability of didymo mat formation increases with higher values of BIO4 and falls with higher values of USGlacier and SegCluesN.

Model Comparisons

The variables included in the final models, together with the proportional contribution of each variable, are shown in Table 3.5. Model fit statistics are shown in Table 3.6.

Two classification thresholds are given for sensitivity and specificity. The first was where the sum of sensitivity and specificity was highest. At this threshold, any increase in sensitivity (by adjusting the classification threshold) led to a larger drop
Figure 3.3: Single variable response plots for temperature seasonality (BIO4) and nitrogen concentration (SegCluesN). Variable responses were calculated for the logistic regression, boosted regression tree (BRT) and Maxent models by holding the value of all other environmental variables in the model at their mean for the training dataset. In all plots, the y-axis is the response and represents the relative probability of didymo mat establishment. Note that the BRT response functions are automatically centered which puts them on a different, though comparable, y-axis scale. Where present, tick marks on the x-axis indicate deciles.

in specificity. The inverse was true also – any increase in specificity led to a larger drop in sensitivity. The second threshold is where specificity was close to 0.5. At this threshold model sensitivity may be interpreted as the proportion of presence locations correctly classified as presences when 50% of the absence locations are misclassified (see Table 3.7).

Predictions for all river segments in New Zealand above stream order three were made using the final model from each modelling algorithm. The mean length of the river segments was 672m ($SD = 642m$). Geographic projections of the predictions were visually examined. The logistic regression prediction (Figure 6) and BRT prediction (Figure 7) were visually more similar to each other than the Maxent prediction was to either of them. In the logistic regression and BRT predictions, Northland, Taranaki, the Auckland region and Stewart Island have all rivers shaded dark green (the lowest risk category for didymo bloom formation). The Maxent prediction, in contrast, has either orange or red river segments in these areas (the second highest and highest risk
Table 3.5: Predictors included in the final Logistic Regression (LogReg), boosted regression tree (BRT) and Maxent models. The proportional contribution of each variable is printed below the predictor name. The predictors are temperature seasonality (BIO4), glacial cover upstream (USGlacier), log-transformed nitrogen concentration (SegCluesN), mean January air temperature (SegJanAirT), proportion of riparian shading (SegRipShad) and days per year of rain greater than 25mm (USDaysRain).

<table>
<thead>
<tr>
<th>Variable Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
</tr>
<tr>
<td>LogReg</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>BRT</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Maxent</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

category, respectively; see Figure 8). A higher proportion of river segments were dark green in the logistic regression model prediction than the BRT model prediction (60% and 53%, respectively). When the dark green and light green categories were pooled, however, there was good agreement between them (91.7% of all dark green or light green river segments were dark green or light green in both prediction maps). All three predictions had a similar proportion of red river segments (between 3.6 and 7.1%) but the locations of these segments in the Maxent prediction differed markedly from the other two. In the Maxent prediction, red river segments were concentrated towards the centre of the North Island. In the logistic regression and BRT prediction, red river segments were concentrated in the centre of the South Island, along the Southern Alps.

There were 182 river segments (0.1%) that were red (at highest risk of didymo bloom formation) in the predictions from all three models and all of them were in the South Island. More specifically, they were in Central Otago and south-west Canterbury on the headwaters of the Oreti, Ahuriri, Hopkins, Shotover and Cass rivers and on several other smaller rivers in the same region.

There were 59,500 river segments (41.7%) that were dark green (relatively unsuitable for didymo bloom formation) in the predictions from all three models. For the South Island, dark green river segments were found in a large part of Southland including Stewart Island and the lower reaches of the Oreti, Mataura and Waiau rivers; Otago in the Taieri, Tokomairiro, Waikouaiti, Kakanui and tributaries of the Clutha River; Canterbury in the Opuha, Hinds, Heathcote, Waipara, Conway and tributaries of the
Table 3.6: Logistic Regression (LogReg), boosted regression tree (BRT) and Maxent model fitting information. The first training dataset (Training 1) contained 39 reliable absences and 51 presences. The second training dataset (Training 2) contained 200 absences and 51 presences. The presence-only dataset contained the 51 training presences only. All models include the number of predictors ($k$) and the proportion of deviance explained ($D^2$). LogReg models have corrected Akaike Information Criterion (AICc) values and the difference in AICc ($\Delta$AICc). BRT models have the learning rate (LR) and number of trees (Trees). Maxent models have training gain (Gain).

<table>
<thead>
<tr>
<th>Logistic Regression</th>
<th>Dataset</th>
<th>Model</th>
<th>$k$</th>
<th>AICc</th>
<th>$\Delta$AICc</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training 1</td>
<td>Global</td>
<td>9</td>
<td>131.6</td>
<td>11.7</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simplified</td>
<td>3</td>
<td>119.9</td>
<td>-</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Training 2</td>
<td>Global</td>
<td>9</td>
<td>240.2</td>
<td>9.3</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simplified</td>
<td>3</td>
<td>230.9</td>
<td>-</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Boosted Regression Trees</th>
<th>Dataset</th>
<th>Model</th>
<th>$k$</th>
<th>LR</th>
<th>Trees</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training 1</td>
<td>Global</td>
<td>9</td>
<td>0.0005</td>
<td>1800</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simplified</td>
<td>4</td>
<td>0.0005</td>
<td>1600</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Training 2</td>
<td>Global</td>
<td>9</td>
<td>0.001</td>
<td>1250</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simplified</td>
<td>3</td>
<td>0.001</td>
<td>1600</td>
<td>0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maxent</th>
<th>Dataset</th>
<th>Model</th>
<th>$k$</th>
<th>Gain</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Presence-only</td>
<td>Global</td>
<td>9</td>
<td>0.352</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simplified</td>
<td>5</td>
<td>0.311</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Table 3.7: Training and Testing sensitivity (Sens), specificity (Spec) and area under the receiver-operating characteristic curve (AUC) for the final logistic regression (LogReg), boosted regression tree (BRT) and Maxent model. The first threshold (A) is at max(sensitivity + specificity). The second threshold (B) was chosen such that specificity is approximately 0.5.

<table>
<thead>
<tr>
<th>Model</th>
<th>Threshold</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>LogReg</td>
<td>(A) 0.19</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>(B) 0.14</td>
<td>0.82</td>
<td>0.50</td>
</tr>
<tr>
<td>BRT</td>
<td>(A) 0.19</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(B) 0.13</td>
<td>0.88</td>
<td>0.46</td>
</tr>
<tr>
<td>Maxent</td>
<td>(A) 0.43</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(B) 0.32</td>
<td>0.90</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Waitaki River; central to northern parts of the West Coast including Arahura, Arnold and the lower Buller River; and Tasman and Marlborough in the Aorere, Moutere, Awatere, Waima and tributaries of the Wairau River. For the North Island, dark green river segments covered most rivers in the Wellington, Gisborne and Auckland regions. Dark green river segments also covered the southern rivers of Northland; coastal parts of the Bay of Plenty; coastal parts of Hawke’s Bay; Waikato in the Awakino, Makau, Piako and Waikato rivers; and the most part of Manawatu-Wanganui including the Akitio, Manawatu, Rangitikei and Wanganui rivers.

Pearson and Spearman correlations were calculated between the predictions. Pearson and Spearman correlation coefficients were highest for the Logistic Regression and BRT model predictions ($r=0.79$, $r_s=0.79$). The Maxent predictions had higher correlations with the Logistic Regression predictions ($r=0.66$, $r_s=0.78$) than the BRT predictions ($r=0.47$, $r_s=0.59$).
Figure 3.4: Presence (red) and absence (green) locations used to train models of the relative suitability of river segments for didymo bloom formation. River segments in the Freshwater Ecosystems of New Zealand (FENZ) network of order three or larger are shown in light blue.
Figure 3.5: Relative suitability of river segments for didymo bloom formation from the final logistic regression model predictions. All river segments in the Freshwater Ecosystems of New Zealand (FENZ) network of order three or larger are colour-coded dark green, light green, orange or red. Dark green indicates a predicted value from the logistic regression model less than 0.14 (this value matches classification threshold 2; the threshold where model specificity approximately equals 0.5). Light green, orange and red colours cover equal intervals in the model predictions between classification threshold 2 and the highest predicted value (0.70). Red indicates highly suitable river segments for didymo bloom formation.
Figure 3.6: Relative suitability of river segments for didymo bloom formation from the final Boosted Regression Tree (BRT) model predictions. All river segments in the Freshwater Ecosystems of New Zealand (FENZ) network of order three or larger are colour-coded dark green, light green, orange or red. Dark green indicates a predicted value from the BRT model less than 0.13 (this value matches classification threshold 2; the threshold where model specificity approximately equals 0.5). Light green, orange and red colours cover equal intervals in the model predictions between classification threshold 2 and the highest predicted value (0.60). Red indicates highly suitable river segments for didymo bloom formation.
Figure 3.7: Relative suitability of river segments for didymo bloom formation from the final Maxent model predictions. All river segments in the Freshwater Ecosystems of New Zealand (FENZ) network of order three or larger are colour-coded dark green, light green, orange or red. Dark green indicates a predicted value from the Maxent model less than 0.32 (this value matches classification threshold 2; the threshold where model specificity approximately equals 0.5). Light green, orange and red colours cover equal intervals in the model predictions between classification threshold 2 and the highest predicted value (0.99). Red indicates highly suitable river segments for didymo bloom formation.
3.4 Discussion

Pfankuch index correlations

The first aim of this chapter was to examine the correlations between environmental layers covering all New Zealand rivers and the Pfankuch index values from the survey described in Chapter 2. This was done in order to find a suitable substitute for the Pfankuch bed stability index value to use in modelling the distribution of didymo. To address this aim, I examined Pearson’s and Spearman’s correlation coefficients for the Pfankuch index value and five environmental layers identified as being relevant to the ecology of didymo from the Freshwater Ecosystems of New Zealand (FENZ) database. The resulting correlation coefficients suggest that all five environmental layers are poor surrogates for Pfankuch bed stability index value. Lake buffering in the upstream catchment (USLake) appears to be an unsuitable substitute because Pfankuch value varies widely when USLake is zero (where there are no lakes upstream) and, critically, most reaches in New Zealand (92.6% of all the river segments) have no lakes upstream. For the other environmental layers, I found the magnitude of all Pearson and Spearman correlation coefficients were less than 0.5. For comparison, Pfankuch’s index is correlated with higher invertebrate species richness with a study of eleven streams yielding \( r_s = -0.75 \) (Death and Winterbourn, 1995) and a second study of 51 streams yielding \( r = -0.36 \) (Townsend et al., 1997). I consider the correlations between Pfankuch index value and the environmental layers used in my work to be quite modest. This means I was unable to find a suitable substitute for Pfankuch index value in the FENZ database and so I did not rely on Pfankuch value to model the distribution of didymo blooms. A comment on the potential for the development of a Pfankuch streambed stability layer is included in the General Discussion.

Modelling the Distribution of Didymo Blooms

The second aim of this chapter was to model the potential distribution of didymo blooms in New Zealand using three distribution modelling algorithms: logistic regression, boosted regression trees and Maxent. By using more than one distribution modelling algorithm, it is possible to examine the influence of algorithm choice on the resulting prediction maps for didymo in New Zealand.

The three performance measures I used to compare the final models to each other and to previous didymo distribution modelling efforts were model sensitivity, model specificity and AUC (area under the receiver operator characteristic curve). To address the above aim of modelling didymo’s potential distribution, a value for model sensitivity
close to one is desirable. A complicating factor is that the value for sensitivity depends on the threshold chosen to calculate it. At first glance, it may appear sensible to choose a threshold of 0.5 but there is no sound ecological basis for this (Liu et al., 2005) and it is somewhat subjective. Instead, I calculated sensitivity and specificity using two, much more objective thresholds: the first such that the sum of sensitivity and specificity are maximised (Threshold A) and the second such that specificity approximately equals 0.5 (Threshold B). As noted in the methods, threshold B helps to answer the question ‘if I allow the model to misclassify 50% of absences, how well does it classify presences?’.

When I allowed the models to misclassify 50% of absences (Threshold B), the three distribution models had testing sensitivities in the range 0.74 to 0.84. This means, on the testing dataset – on which the models had not been trained – the models correctly predicted 74 to 84% of the didymo presence locations. In a study modelling the potential distribution of didymo in the USA, didymo presence locations were correctly predicted up to 88% of the time (Kumar et al., 2009) while maintaining values for specificity between 70% and 80%. The difference in performance between this study and mine may be due to sample size (Kumar et al. used six times more presence locations than in my study), the particular environmental layers included in the modelling, the choice of smallest study unit (pixels versus river segments), the particular aspects of the study region (US versus New Zealand) or other factors. Comparisons were not made with the two didymo modelling studies in New Zealand, Kilroy et al. (2007) and Kilroy et al. (2008), because they do not report sensitivity or specificity.

With regard to AUC, the boosted regression tree model was the highest performing for the training dataset and the logistic regression model was the highest performing for the testing dataset. There were, however, only small differences in training and testing AUC values between models. I have not compared the AUC values in this study to other distribution modelling efforts because it is not appropriate to make comparisons between models built from different datasets using this measure (Yackulic et al., 2013). There is also no AUC value above which a model may be considered ‘very accurate’. A low value of AUC can occur when a species is limited by dispersal which is likely for newly-invading organisms (Peterson et al., 2008).

Response functions are a common way of examining model behaviour in species distribution modelling (Guisan and Zimmermann, 2000; Elith and Leathwick, 2009). Prior to fitting the distribution models, I recorded the expected direction (and where possible the shape) of the response function for each of the nine environmental variables (see Table 3.2) in relation to didymo growth. The two most important response functions to examine are for the two environmental variables retained by all the fitted models. These variables were temperature seasonality (BIO4) and nitrogen concentra-
tion (SegCluesN; see Figure 3.3). So how well did the observed response functions from the distribution models match what was expected?

**BIO4:** Higher relative suitabilities of didymo mat formation were expected with increasing temperature seasonality (BIO4). The response plots resulting from the logistic regression, BRT and Maxent models showed an increasing trend except for values of BIO4 below 3000 in the Maxent model. It is unclear why this trend reversal occurred. Perhaps it is an artifact of having few data points in the training dataset with low temperature seasonalities (only 1 out of 51 presence locations had BIO4 less than 3000). Although there are few river segments with BIO4 less than 3000 across the country (just 3.6%), they are found in the areas where the Maxent model predictions differ from the BRT and Logistic Regression model predictions. Notably, Northland has a high proportion of river segments with low values of BIO4 and the Maxent prediction has many more red and orange rivers segments there compared to the BRT and logistic regression predictions. For Northland, and the small number of other river segments with BIO4 less than 3000, I consider the predictions from the BRT and logistic regression models to be more realistic than the predictions from the Maxent model.

**SegCluesN:** Lower relative suitabilities of didymo mat formation were expected with increasing nitrogen concentrations (SegCluesN). This trend was seen in all response plots with a tendency for low (close to 0) relative probabilities of didymo occurrence above 3.5 parts per billion of nitrate-nitrogen (corresponding to 1.5 on the log x+1 transformed x-axis in Figure 3.3).

In addition to BIO4 and SegCluesN, the final logistic regression and Maxent models included the proportion of glacial cover in the upstream catchment (USGlacier). Lower probabilities of didymo mat formation were expected with increasing upstream glacial cover. This relationship was expected due to the importance of light for didymo growth (Cullis et al., 2012) and the tendency for glaciers to produce light-attenuating glacial flour during the warmer months of the year. The response plots for USGlacier all had a negative trend, matching the expected relationship.

Overall, the close match between expected and observed response functions is a good indication that the distribution models give useful predictions for didymo mat establishment in New Zealand.

The predictions of didymo bloom formation may have differed if I chose to include the North Island survey data in the training of the distribution models. If the North Island data was included in the model training, I would expect the sensitivity of all the distribution models to fall. This means the predictions of didymo bloom formation would be less reliable. In future investigations, it may be appropriate to compare didymo distribution predictions with and without North Island survey data.
Three environmental variables were not retained in any of the models because their contribution (the proportion of deviance explained by the variable) was small. They were DAYS_BTW_F (mean days between 3x median flow), USLake (lake buffering in upstream catchment) and USDaysRain (days per year with rain > 25mm in the upstream catchment). All three relate to flow variability. It is possible that at smaller scales - catchment scale or below - these flow variables are useful predictors of the relative suitability of river segments for didymo bloom formation but, at a New Zealand scale, temperature variability (BIO4 in this case) appears to dominate.

**Temperature Seasonality**

I found that in all three models, temperature seasonality (BIO4) was the most important predictor of didymo mat occurrence. It contributed almost three-quarters of the explanatory power of the final logistic regression model and close to 60% in both the BRT and Maxent models. This predictor represents the variation in temperature during the year and it is calculated by multiplying the standard deviation in monthly temperatures by 100 (Hijmans *et al.*, 2005).

As temperature seasonality increased, there was an increase in the predicted suitability of river segments for didymo bloom formation. This response was expected because greater temperature seasonality increases the probability that, at some time during the year, river water temperature is within the range permitting didymo mat establishment. That ranges is likely to be below 20°C because Kilroy *et al.* (2007) found a distinct cut-off in didymo occurrence with summer temperatures above this temperature. But, in areas with summer temperatures higher than 20°C, didymo establishment may still be possible in winter (if temperature seasonality is high).

**Model Comparisons**

The third aim of this study was to create and compare geographic projections (maps) of the distribution models. The visual similarity between the maps suggests that, given the modelling dataset chosen, there is some flexibility in the choice of distribution modelling algorithm.

The predictions of didymo mat growth of the current study are broadly in agreement with the rapid assessment maps produced by Kilroy *et al.* (2008). They found, as did I, that there are far more river segments in the South Island in the highest risk category compared to the North Island; that Otago and Canterbury are generally in higher risk categories than Tasman or Southland; and that red and orange river segments are concentrated at the centre of the North Island. At finer scales, the differences in
prediction maps are more pronounced. For example, the Wairau, Awatere and Waima rivers in Marlborough are predominantly dark green in the prediction maps of this study but appear orange or red in the suitability map of Kilroy et al. (2008).

The prediction maps of the current study are also broadly similar to those produced by Kilroy et al. (2007). Although the colour schemes differ, their results showed generally higher didymo mat thicknesses in the South Island and the thickest mats in regulated rivers.

The main conclusions I can draw from this chapter are threefold. First, Pfankuch bed stability may be a useful environmental layer but there is no suitable substitute available at present. Second, the choice of distribution modelling algorithm for modelling the distribution of didymo in New Zealand has a greater influence on the resulting predictions at the river segment scale than at the regional scale. Finally, the distribution of didymo mats in New Zealand on a regional scale appears to be heavily dependent on temperature seasonality.
Chapter 4

General Discussion

Research on didymo blooms has been a focus for many ecologists since the species’ rapid spread in several countries. My aim in this thesis was to provide more information on the Why and the Where: ‘What are the conditions required for didymo blooms to form?’ (Chapter 2) and ‘Where are those conditions found in New Zealand?’ (Chapter 3). To address the Why, I examined the relationship between streambed stability and didymo blooms in a field survey. Pfankuch streambed stability was estimated at forty river sites along with two measures of didymo bloom magnitude – percentage cover and standing crop. I found that the Pfankuch value was a stronger predictor (in terms of model likelihoods) of percentage cover and standing crop compared to spot measurements of dissolved reactive phosphorus concentration and turbidity. To address the Where, I modelled the relative suitability of river segments in New Zealand for didymo bloom formation. The models were built on a database of occurrence locations for didymo blooms and a suite of environmental layers covering the New Zealand river network. I used three distribution modelling algorithms, namely logistic regression, boosted regression trees and Maxent. I found that the environmental layer representing temperature seasonality had the highest proportional contribution to the three final models and that the reaches with the highest risk of didymo blooms were in Otago, Canterbury and the West Coast regions of the South Island of New Zealand.

4.1 Study Limitations

An important result of my didymo bloom distribution modelling was a set of prediction maps showing the relative suitability of river segments for didymo blooms. These predictions were consistent with the current knowledge of the distribution of didymo blooms on a large spatial scale, but there were discrepancies with previous distribution
models when examined in detail. For example, the Didymo Samples Database indicates that didymo formed blooms in the Buller River in 2005 (MAF Biosecurity NZ, 2012) but all segments of this river show up green (low risk) in the logistic regression and boosted regression tree models and mostly green in the Maxent model. An explanation for this contradiction is that the single record in the Buller River during 2005 was excluded from the distribution modelling because it had no site coordinates attached to it. Site coordinates were essential to connect occurrence records in the Database to river segments in the environmental layers. The more general point is that my distribution models for didymo only ‘know’ (are training on) a subset of all recorded didymo bloom occurrence locations. Nevertheless, I would argue that the occurrence data used to train the models are representative of the conditions under which didymo blooms form. There will be, however, opportunities to improve on my didymo distribution models as didymo spreads. New records are regularly being added to the Didymo Samples Database and, increasingly, information about the characteristics of didymo blooms is being recorded in it (MAF Biosecurity NZ, 2012). Two more aspects of my didymo distribution modelling which could be improved are the suite of environmental layers used and the method of reducing sampling bias in the occurrence locations.

First, environmental layers matter in species distribution modelling (Peterson and Nakazawa, 2008). I used environmental layers from the Freshwater Ecosystems of New Zealand (FENZ) database (not to be confused with the didymo samples database). The FENZ database was constructed from geographic, geological and meteorological data collected over several decades (Snelder and Biggs, 2002). The values of environmental layers in many parts of the country will have changed since then. Riparian shading and nitrogen concentrations, for example, have changed with land use change along rivers throughout the country (Moller et al., 2008). Also, the environmental layers are all models of the variables they represent. None of the layers contain physical measurements at every river segment in New Zealand. Rather, interpolation and extrapolation of existing data have been used to assign values to every segment. In recognition of this limitation, I used performance measures and cross-validation (splitting the occurrence locations into training and testing datasets) to check that the models produce sensible predictions.

Second, in species distribution modelling there is an assumption that the sampling locations are taken completely at random from the geographic area where the species predictions are made. In practice, however, there is considerable bias in the selection of sampling sites. One form of bias is to sample only at or close to roads. If all samples are collected close to roads then the species distribution may reflect the conditions around roads rather than the actual distribution of the species (Phillips et al., 2009).
Several methods have been proposed to deal with this problem. An obvious solution is to collect data from undersampled areas or remove data from oversampled areas (sensu Dudik et al., 2006). I applied the second method by thinning clusters of sites within 10km down to a single site. There are other solutions to this problem including the suggestion by Phillips et al. (2009) that the method of sampling background points (part of the Maxent modelling process) is allowed to have a similar bias as the sampling of occurrence locations. For example, a set of occurrence locations are collected from within 1km of road sides within a large geographic area. The model would normally take a random sample of pixels from the full geographic area to approximate the mean environmental conditions. This time though, these background samples are also taken from within 1km of road sides. This change is intended to bias the environmental conditions in such a way that the sampling bias no longer sways the resulting distribution towards road sides. This approach results in model performance improvements (Phillips et al., 2009).

A further limitation of my species distribution models for didymo is that the predictions for the North Island are based only on didymo occurrence data from the South Island. That is, no presence and absence locations from the North Island were used to train the models but I have extrapolated the predictions to this region. This extrapolation involved as element of risk because the model predictions may not be sensible outside the range of values the model was trained on. Take temperature seasonality (the standard deviation of temperature multiplied by 100) to emphasise this point. The range of training values for temperature seasonality was 2948 to 4575 which does not include the lower values found on the North Island (ranging from 2724 to 3760).

One approach to characterising the ‘extrapolation uncertainty’ is to calculate a dissimilarity index between the values of the environmental layers at each reach and the range of values in the training dataset. This index is zero when all variables take values within the training range and positive when at least one variable is outside the training range. This approach was taken by Kilroy et al. (2007). They found that, in general, predictions for the North Island were less reliable than those for the South Island.

I have avoided applying a dissimilarity index to my own data because the range of training values for the variables included in the final models overlapped considerably with the values in New Zealand. Also, this approach to calculating model reliability does not account for the respective contributions made by each variable to the model predictions. In my case, being outside the range of temperature seasonalities is much more of a problem than being outside the range of USDaysRain because of a dramatic difference in variable contributions (40% versus 3% in the final Maxent model).
4.2 Further Research Needs

There are numerous questions which arose during this work, many of which make good candidates for further research.

In Chapter 2, I examined the relationship between Pfankuch streambed stability index value and didymo biomass. I found that Pfankuch bed stability is a more informative predictor of didymo biomass than nutrient concentrations. It remains unclear whether there is a threshold of streambed stabilities, above which visible didymo colonies are able to form.

An important step, I argue, in improving the modelling of didymo bloom formation is to develop an environmental layer representing Pfankuch streambed stability. This could be achieved by collating Pfankuch bed stability index values from existing research and collecting additional data in poorly-covered areas. Such a dataset could be used to model Pfankuch bed stability in all New Zealand river segments based on existing environmental layers. In this context, I should caution that the environmental layers in the FENZ database may not be suitable for this task because of a poor match between the components of bed stability and the available layers. There are, however, other databases containing environmental layers which may be more suitable.

The purpose of using three different distribution modelling algorithms to model the distribution of didymo mats was to compare the predictions between models. This comparison allowed one element of the species distribution modelling process to be examined. Another element amenable to manipulation is the suite of environmental layers. I chose a list of nine abiotic environmental layers which had a plausible link to didymo mat establishment from the available layers in the FENZ and WorldClim databases. A sensible next step would be to add dispersal limitation variables – such as a measure of human accessibility – to the list of environmental layers and compare model predictions. This approach was taken by Compton et al. (2012) for modelling the distribution of four introduced macrophytes in New Zealand lakes using boosted regression trees. Assuming that humans are important vectors for spreading these macrophyte species, they added distance to the nearest highway, road density and human population density near the lakes to the list of environmental layers. These variables turned out to be important predictors of the distribution of the four macrophytes. Similarly, humans are considered important vectors for the spread of didymo (Kilroy and Unwin, 2011). It is likely that adding similar environmental layers to future distribution models would be useful in extending our understanding of didymo’s distribution.
4.3 Concluding note

In writing this thesis, I have gained an understanding of the difficulties present in predicting where didymo blooms are likely to occur in New Zealand rivers. The task is simple in a post-hoc manner because when do we see didymo blooms occurring, then those locations are obviously in the high risk category. By applying species distribution modelling, I have attempted to make predictions about areas where didymo blooms have never been observed. In the end, the true test of the distribution models will be to wait and see where didymo spreads and where it forms these nuisance blooms.
References


65


LINZ (2013). Land Information New Zealand Data Service.


Appendix A

Didymo survey data

Data collected to examine the relationship between streambed stability and didymo establishment (Chapter 2)
<table>
<thead>
<tr>
<th>No.</th>
<th>Site</th>
<th>Waterway</th>
<th>Latitude WGS84</th>
<th>Longitude WGS84</th>
<th>Stream</th>
<th>SCI</th>
<th>Cover (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AHU</td>
<td>Ahuriri</td>
<td>-44.46995</td>
<td>169.98866</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>APA</td>
<td>Aparima</td>
<td>-46.03890</td>
<td>168.11731</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>ARA</td>
<td>Arahura</td>
<td>-42.70254</td>
<td>171.05954</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>ARR</td>
<td>Arrow</td>
<td>-44.93689</td>
<td>168.83246</td>
<td>4</td>
<td>109</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>ASH</td>
<td>Ashburton Sth</td>
<td>-43.74190</td>
<td>171.52863</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>CASS</td>
<td>Cass</td>
<td>-43.87241</td>
<td>169.46298</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>CLU</td>
<td>Clutha</td>
<td>-44.68123</td>
<td>169.18942</td>
<td>7</td>
<td>408</td>
<td>62</td>
</tr>
<tr>
<td>8</td>
<td>DAR</td>
<td>Dart</td>
<td>-44.77605</td>
<td>168.32520</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>GREY</td>
<td>Grey</td>
<td>-42.44738</td>
<td>171.25620</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>HAAST</td>
<td>Haast</td>
<td>-44.01844</td>
<td>169.37286</td>
<td>5</td>
<td>46</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>HAK</td>
<td>Hakatamea</td>
<td>-44.72586</td>
<td>170.49045</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>HORO</td>
<td>Hororata</td>
<td>-43.54380</td>
<td>171.95256</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>KAK</td>
<td>Kakanui</td>
<td>-45.11702</td>
<td>170.82790</td>
<td>6</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>LIN</td>
<td>Lindis</td>
<td>-44.88046</td>
<td>169.35378</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>LWAI</td>
<td>Lower Waiau</td>
<td>-46.03000</td>
<td>167.71274</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>MAK</td>
<td>Makarora</td>
<td>-44.15898</td>
<td>169.29089</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>MARA</td>
<td>Mararoa</td>
<td>-45.60677</td>
<td>167.70193</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>MAT</td>
<td>Mataura</td>
<td>-46.18257</td>
<td>168.87930</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>MIK</td>
<td>Mikonui</td>
<td>-42.90734</td>
<td>170.76320</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>MTKK</td>
<td>Matukituki</td>
<td>-44.56225</td>
<td>168.90213</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>NEV</td>
<td>Nevis</td>
<td>-45.17048</td>
<td>168.99602</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>OHAU</td>
<td>Ohau</td>
<td>-44.26540</td>
<td>169.98340</td>
<td>6</td>
<td>289</td>
<td>97</td>
</tr>
<tr>
<td>23</td>
<td>OPI</td>
<td>Opahi</td>
<td>-44.09707</td>
<td>170.83453</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>OPUHA</td>
<td>Opua</td>
<td>-44.08110</td>
<td>170.97900</td>
<td>6</td>
<td>56</td>
<td>25</td>
</tr>
<tr>
<td>25</td>
<td>ORARI</td>
<td>Orari</td>
<td>-44.04436</td>
<td>171.26895</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>ORE</td>
<td>Oreti</td>
<td>-45.65908</td>
<td>168.22096</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>OTE</td>
<td>Otematata</td>
<td>-44.60552</td>
<td>170.20233</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>PAR</td>
<td>Paringa</td>
<td>-43.71165</td>
<td>169.49020</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>POM</td>
<td>Pomahaka</td>
<td>-46.05433</td>
<td>169.24063</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>PRIN</td>
<td>Princhester</td>
<td>-45.54589</td>
<td>167.93307</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>RAK</td>
<td>Rakaia</td>
<td>-43.71189</td>
<td>171.98158</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>RANG</td>
<td>Rangitata</td>
<td>-43.97546</td>
<td>171.30188</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>SHO</td>
<td>Shotover</td>
<td>-44.99994</td>
<td>168.75745</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>TEK</td>
<td>Tekapo</td>
<td>-44.01001</td>
<td>170.47737</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>TOT</td>
<td>Totara</td>
<td>-42.88396</td>
<td>170.84569</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36</td>
<td>TWI</td>
<td>Twizel</td>
<td>-44.25523</td>
<td>170.11039</td>
<td>5</td>
<td>43</td>
<td>17</td>
</tr>
<tr>
<td>37</td>
<td>WAIH</td>
<td>Waihao</td>
<td>-44.79512</td>
<td>171.09583</td>
<td>5</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>38</td>
<td>WAIM</td>
<td>Waimakariri</td>
<td>-43.44037</td>
<td>172.19533</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>WAIR</td>
<td>Wairaki</td>
<td>-45.93448</td>
<td>167.70479</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>WAIT</td>
<td>Waitaki</td>
<td>-44.72866</td>
<td>170.47792</td>
<td>8</td>
<td>248</td>
<td>65</td>
</tr>
<tr>
<td>Site</td>
<td>Depth (cm)</td>
<td>Velocity ($cms^{-1}$)</td>
<td>Temp ($^\circ$C)</td>
<td>Pfankuch Index ($\mugL^{-1}$)</td>
<td>DRP NNN ($\mugL^{-1}$)</td>
<td>Turb (NTU)</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>------------------------</td>
<td>------------------</td>
<td>-------------------------------</td>
<td>----------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>AHU</td>
<td>28</td>
<td>1.13</td>
<td>10.9</td>
<td>54</td>
<td>3</td>
<td>30</td>
<td>1.15</td>
</tr>
<tr>
<td>APA</td>
<td>36</td>
<td>0.89</td>
<td>14.2</td>
<td>51</td>
<td>5</td>
<td>280</td>
<td>0.45</td>
</tr>
<tr>
<td>ARA</td>
<td>42</td>
<td>0.81</td>
<td>12.1</td>
<td>43</td>
<td>2.5</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>ARR</td>
<td>32</td>
<td>1.31</td>
<td>6.8</td>
<td>44</td>
<td>2</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>ASH</td>
<td>30</td>
<td>1.06</td>
<td>12.4</td>
<td>36</td>
<td>2</td>
<td>80</td>
<td>0.5</td>
</tr>
<tr>
<td>CASS</td>
<td>30</td>
<td>0.81</td>
<td>13.2</td>
<td>47</td>
<td>2.5</td>
<td>20</td>
<td>2.55</td>
</tr>
<tr>
<td>CLU</td>
<td>48</td>
<td>0.57</td>
<td>13.1</td>
<td>22</td>
<td>0.5</td>
<td>30</td>
<td>0.7</td>
</tr>
<tr>
<td>DAR</td>
<td>30</td>
<td>0.72</td>
<td>12.6</td>
<td>50</td>
<td>1.5</td>
<td>15</td>
<td>5.55</td>
</tr>
<tr>
<td>GREY</td>
<td>22</td>
<td>0.61</td>
<td>14.8</td>
<td>36</td>
<td>2.5</td>
<td>80</td>
<td>0.25</td>
</tr>
<tr>
<td>HAAST</td>
<td>44</td>
<td>0.47</td>
<td>9.2</td>
<td>45</td>
<td>1.5</td>
<td>10</td>
<td>0.2</td>
</tr>
<tr>
<td>HAK</td>
<td>22</td>
<td>0.86</td>
<td>17.4</td>
<td>52</td>
<td>6</td>
<td>15</td>
<td>0.2</td>
</tr>
<tr>
<td>HORO</td>
<td>24</td>
<td>0.79</td>
<td>13.3</td>
<td>52</td>
<td>6</td>
<td>865</td>
<td>0.2</td>
</tr>
<tr>
<td>KAK</td>
<td>24</td>
<td>0.72</td>
<td>16.1</td>
<td>50</td>
<td>3</td>
<td>60</td>
<td>0.25</td>
</tr>
<tr>
<td>LIN</td>
<td>34</td>
<td>1.03</td>
<td>15</td>
<td>41</td>
<td>2</td>
<td>0</td>
<td>0.35</td>
</tr>
<tr>
<td>LWAI</td>
<td>38</td>
<td>0.63</td>
<td>13.5</td>
<td>19</td>
<td>1</td>
<td>175</td>
<td>0.6</td>
</tr>
<tr>
<td>MAK</td>
<td>36</td>
<td>0.76</td>
<td>10.4</td>
<td>37</td>
<td>2</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>MERA</td>
<td>40</td>
<td>0.70</td>
<td>10.4</td>
<td>41</td>
<td>8.5</td>
<td>315</td>
<td>0.3</td>
</tr>
<tr>
<td>MAT</td>
<td>52</td>
<td>0.36</td>
<td>11.4</td>
<td>51</td>
<td>5.5</td>
<td>330</td>
<td>0.95</td>
</tr>
<tr>
<td>MIK</td>
<td>44</td>
<td>0.62</td>
<td>14.1</td>
<td>37</td>
<td>1</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>MTKK</td>
<td>30</td>
<td>0.58</td>
<td>14.8</td>
<td>53</td>
<td>1</td>
<td>20</td>
<td>0.55</td>
</tr>
<tr>
<td>NEV</td>
<td>60</td>
<td>1.05</td>
<td>8.2</td>
<td>24</td>
<td>2.5</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>OHAU</td>
<td>30</td>
<td>0.84</td>
<td>14.4</td>
<td>21</td>
<td>0.5</td>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td>OPI</td>
<td>35</td>
<td>1.34</td>
<td>16.5</td>
<td>35</td>
<td>5</td>
<td>345</td>
<td>0.25</td>
</tr>
<tr>
<td>OPUHA</td>
<td>22</td>
<td>0.71</td>
<td>17.4</td>
<td>32</td>
<td>1.5</td>
<td>230</td>
<td>0.45</td>
</tr>
<tr>
<td>ORARI</td>
<td>38</td>
<td>1.07</td>
<td>17.7</td>
<td>50</td>
<td>3.5</td>
<td>15</td>
<td>1.9</td>
</tr>
<tr>
<td>ORE</td>
<td>26</td>
<td>0.51</td>
<td>13.3</td>
<td>44</td>
<td>3</td>
<td>55</td>
<td>0.2</td>
</tr>
<tr>
<td>OTE</td>
<td>24</td>
<td>0.86</td>
<td>11.3</td>
<td>54</td>
<td>2.5</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>PAR</td>
<td>60</td>
<td>0.32</td>
<td>12.3</td>
<td>18</td>
<td>2</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>POM</td>
<td>42</td>
<td>0.90</td>
<td>14.9</td>
<td>37</td>
<td>4</td>
<td>215</td>
<td>1.25</td>
</tr>
<tr>
<td>PRIN</td>
<td>18</td>
<td>0.78</td>
<td>12.6</td>
<td>51</td>
<td>3.5</td>
<td>95</td>
<td>0.35</td>
</tr>
<tr>
<td>RAK</td>
<td>26</td>
<td>0.98</td>
<td>17.3</td>
<td>49</td>
<td>2</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>RANG</td>
<td>32</td>
<td>0.69</td>
<td>14.7</td>
<td>28</td>
<td>1.5</td>
<td>30</td>
<td>0.75</td>
</tr>
<tr>
<td>SHO</td>
<td>24</td>
<td>0.94</td>
<td>10.9</td>
<td>50</td>
<td>2</td>
<td>10</td>
<td>4.75</td>
</tr>
<tr>
<td>TEK</td>
<td>40</td>
<td>1.01</td>
<td>10.1</td>
<td>27</td>
<td>1</td>
<td>20</td>
<td>1.2</td>
</tr>
<tr>
<td>TOT</td>
<td>24</td>
<td>0.84</td>
<td>18</td>
<td>48</td>
<td>1</td>
<td>20</td>
<td>0.25</td>
</tr>
<tr>
<td>TWI</td>
<td>46</td>
<td>0.75</td>
<td>14.7</td>
<td>46</td>
<td>2</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>WAIIH</td>
<td>24</td>
<td>0.77</td>
<td>18.6</td>
<td>48</td>
<td>2</td>
<td>50</td>
<td>0.35</td>
</tr>
<tr>
<td>WAIM</td>
<td>42</td>
<td>1.11</td>
<td>17.6</td>
<td>53</td>
<td>2</td>
<td>60</td>
<td>2.4</td>
</tr>
<tr>
<td>WAIR</td>
<td>34</td>
<td>0.82</td>
<td>12.7</td>
<td>49</td>
<td>3</td>
<td>30</td>
<td>0.6</td>
</tr>
<tr>
<td>WAIT</td>
<td>26</td>
<td>0.67</td>
<td>14.8</td>
<td>33</td>
<td>5.5</td>
<td>5</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Appendix B

Pfankuch index form

Form used to determine Pfankuch streambed stability. The bottom component of the Pfankuch index is called 'Stream Bed'.
<table>
<thead>
<tr>
<th>UPPER BANKS</th>
<th>EXCELLENT</th>
<th>GOOD</th>
<th>FAIR</th>
<th>POOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landform slope</td>
<td>Bank slope gradient &lt;1%</td>
<td>Bank slope gradient 1% - 4%</td>
<td>Bank slope gradient 4% - 6%</td>
<td>Bank slope gradient &gt;6%</td>
</tr>
<tr>
<td>Mass-wasting (existing or potential)</td>
<td>No evidence of post or any potential for future mass-wasting into channel.</td>
<td>Intermittent and/ or very small, mostly heaved over. Low is no potential.</td>
<td>Moderate frequency and size, with some low spots edged by water during high flow.</td>
<td>Frequent or large, causing sediment OR imminent danger of some.</td>
</tr>
<tr>
<td>Debris jams potential (floatable objects)</td>
<td>Essentially, are sent from immediate channel area.</td>
<td>Present but mostly small twigs and limbs.</td>
<td>Present, volume and size are both increasing.</td>
<td>Moderate to heavy amounts, mainly larger sizes.</td>
</tr>
<tr>
<td>Vegetative bank protection</td>
<td>&gt;90% plant density. Vigor and variety suggest a deep, dense, soil binding root mass.</td>
<td>70-90% plant density. Fewer species or lower vigor suggest a less dense or deep root mass.</td>
<td>50-70% density. Lower vigor and species from a somewhat shallow and discontinuous root mass.</td>
<td>&lt;50% density and vigor and species indicate discontinuous and shallow root mass.</td>
</tr>
<tr>
<td>LOWER BANKS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bank rock content</td>
<td>65% with large, angular boulders 30cm in number.</td>
<td>40 to 65%, mostly small boulders to cobbles 15-30cm.</td>
<td>20 to 40%, with most in the 7.5-15 cm diameter class.</td>
<td>&lt;20% rock fragments of gravel sizes, 2.5-7.5 cm or less.</td>
</tr>
<tr>
<td>Obstructions (Rock debris Sediment traps)</td>
<td>Rocks and old logs firmly embedded. Flow pattern is not outwearing or deposition.</td>
<td>Some present, causing erosion of current and minor pool filling. Obstructions and deflectors move with high water causing bank cutting and filling of pools.</td>
<td>Moderately frequent, unstable obstructions and deflectors move with high water causing bank cutting and filling of pools.</td>
<td>Frequent obstructions and deflectors cause bank erosion. Sediment traps full channel migration occurring.</td>
</tr>
<tr>
<td>Undercutting</td>
<td>Little or none evident. Intermittent low banks &lt;10cm high.</td>
<td>Some, intermittently at subsurfaces and channel banks. Flow banks &lt;10cm.</td>
<td>Significant. Cuts 15-30cm high. Pool material and slumping evident.</td>
<td>Almost continuous cuts, some &lt;50cm high. Failure of overhangs</td>
</tr>
<tr>
<td>Deposition</td>
<td>Little or no enlargement of channel or points bars.</td>
<td>Some new increase in bar formation, mostly from coarse gravels.</td>
<td>Moderate deposition of new gravel and coarse sand on old and some new bars.</td>
<td>Extensive deposition of gravel and silt, particles, Accreted bars.</td>
</tr>
<tr>
<td>STREAM BED</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rock angularity</td>
<td>Sharp edges and corners, plane surfaces roughened.</td>
<td>Rounded corners and edges, smooth and flat.</td>
<td>Corners and edges well rounded in two dimensions.</td>
<td>Well rounded in all dimensions.</td>
</tr>
<tr>
<td>Brightness</td>
<td>Surfaces dull, darkened or stained. Not &quot;bright.&quot;</td>
<td>Mostly dull, but may have up to 35% bright surfaces.</td>
<td>Mixtures, 50-60% dull and bright i.e. 35-45%</td>
<td>Predominantly bright, 65% exposed surfaces.</td>
</tr>
<tr>
<td>Consolidation or particle packing</td>
<td>Assorted sizes tightly packed and over-lapping.</td>
<td>Moderately packed with some over-lapping.</td>
<td>Mostly a loose assortment with some apparent overlay.</td>
<td>No packing evident. Loose, easily moved.</td>
</tr>
<tr>
<td>Bottom surface distribution &amp; stable</td>
<td>No change in sizes evident, Stable materials 80-100%</td>
<td>Distribution shift slight, Stable materials 50-80%</td>
<td>moderate change in sizes, Stable materials 20-50%</td>
<td>Marked change. Stable materials 0-20%</td>
</tr>
<tr>
<td>Scouring and deposition</td>
<td>45% of the bottom affected by scouring and deposition.</td>
<td>50-60% affected. Scour at constrictions and where deep. Pool deposition.</td>
<td>50-65% affected. Deposition and scour at obstructions, constrictions, and bends.</td>
<td>&gt;50% of bed in a state of flux or change, nearly year round.</td>
</tr>
<tr>
<td>Clinging a quiet vegetation (moss and algae)</td>
<td>Abundant, growth largely moss, dark green, perennials, in swift water too.</td>
<td>Common, Algae forms in low velocity and pool areas. Moss and swifter waters.</td>
<td>Present but spotty, mostly in backwater areas. Seasonal blooms.</td>
<td>Potential types scarse or absent. Yellow-green, short term bloom present.</td>
</tr>
</tbody>
</table>

COLUMN TOTALS

Reach score: <38 = Excellent, 39-76 = Good, 77-114 = Fair, 115+ = Poor
Appendix C

Violin plots

Violin plots of the environmental variables used to model the relative suitability of river segments for didymo bloom formation (Chapter 3)
Figure C.1: Violin plots of the proportion of riparian shading (SegRipShad), glacial cover in the upstream catchment (USGlacier) and log-transformed nitrogen concentration (SegCluesN). Separate violin plots are shown for the training and testing datasets for comparison against values at all didymo presence (N=125) and absence (N=975) locations surveyed in New Zealand (All NZ). There were two training datasets: Training 1 contained 39 reliable absences and 51 training presences. Training 2 contained 200 absences and the same set of 51 training presences. The testing dataset contained 50 presences and 50 absences.
Figure C.2: Violin plots of mean January air temperature (SegJanAirT), temperature seasonality (BIO4) and bed sediment classes weighted by proportional cover (ReachSed). Separate violin plots are shown for the training and testing datasets for comparison against values at all didymo presence (N=125) and absence (N=975) locations surveyed in New Zealand (All NZ). There were two training datasets: Training 1 contained 39 reliable absences and 51 training presences. Training 2 contained 200 absences and the same set of 51 training presences. The testing dataset contained 50 presences and 50 absences.
Figure C.3: Violin plots of lake buffering in the upstream catchment (USLake), days per year with rain greater than 25mm in the upstream catchment (USDaysRain) and log-transformed mean number of days between floods 3x median flow (DAYS_BTW_F). Separate violin plots are shown for the training and testing datasets for comparison against values at all didymo presence (N=125) and absence (N=975) locations surveyed in New Zealand (All NZ). There were two training datasets: Training 1 contained 39 reliable absences and 51 training presences. Training 2 contained 200 absences and the same set of 51 training presences. The testing dataset contained 50 presences and 50 absences.