Spatial data for modelling and management of freshwater ecosystems

Yung En Chee* and Jane Elith

Australian Centre of Excellence for Risk Analysis, School of Botany, University of Melbourne, Parkville, VIC 3010, Australia

(Received 1 December 2011; final version received 25 July 2012)

Fluvial habitats are inherently variable. They are shaped by flow magnitude, frequency, timing and duration, by the effects of upstream and downstream features along flow paths and by bioclimatic processes and human activities in upstream contributing catchments. Managing freshwater ecosystems requires tools and data that effectively account for these multi-scale processes. We tackle these challenges in this analysis of the distribution of 17 native and alien fish species in south-eastern Australia. A fine-scale, stream-link-based GIS database comprising an extensive set of ecologically meaningful attributes at multiple scales was developed to characterise the multidimensional environmental space of freshwater biota. This article describes the methods and data required to construct such a database. Boosted regression tree models were employed to analyse relationships between species and 20 candidate environmental predictors. For some species, competitors/predators were also included as predictors. Models were evaluated from several viewpoints: the ecological plausibility and intuition arising from them, their ability to predict to river links within the training area and for 11 species for which data were sufficient, their ability to predict to an adjacent but geographically distinct region. Despite modest environmental contrasts in the study area, these data and species distribution models (SDMs) produced predictions with useful predictive ability and discriminatory power. Critically, predictors of distribution identified as important for the various species modelled were ecologically interpretable. Several—but not all—of the models tested for transferability also predicted distributions reasonably well in the adjacent region. The GIS stream database and SDMs have immediate applications, but also provide a valuable foundation for developing more sophisticated tools for management and conservation in Australian freshwater environments.

Keywords: multi-scale; ecologically meaningful attributes; fish species distribution models; boosted regression trees; prediction; transferability

Introduction

The freshwater ecosystems of south-eastern Australia are becoming increasingly degraded via anthropogenic pressures. Direct pressures include dam construction, flow regulation, stream channelisation, de-snagging, draining of wetlands and construction of levees. Indirect pressures include native vegetation clearing, agricultural development and the consequences of erosion, sedimentation, nutrient run-off and alien species introductions (CES 2008). The profound changes in our riverine systems and the challenges of climate change have increased the impetus for conserving freshwater-dependent biota and redressing the...
degradation. In this context, predictive models of relationships between environment and the distribution of biota are valuable tools.

Species distribution models (SDMs) have wide-ranging applications including: quantifying species–habitat relationships; identifying unsurveyed sites of high potential occurrence for rare species (Engler et al. 2004); supporting species recovery and reintroduction plans (Steel et al. 2004, Martinez-Meyer et al. 2006); contributing inputs for conservation planning (Moilanen et al. 2008, Leathwick et al. 2008b); assessing impacts of climate and land-use changes on species distribution (Thuiller et al. 2008); predicting species invasion (Hartley et al. 2006) and designing cost-effective surveillance for invasive species (Hauser and McCarthy 2009).

SDMs have been widely implemented in terrestrial and marine environments, but more sparingly in freshwaters. This might reflect the availability of relevant data for modelling, but also perhaps the complexity of the modelling task. Fish occurrence and abundance patterns typically exhibit complex, non-linear relationships to habitat heterogeneity and biotic interactions (Olden and Jackson 2002). Habitats in fluvial systems are heterogeneous and nuanced, shaped as they are by the interactions of flow magnitude, frequency, timing and duration within the geomorphic templates. Additional complexity is overlaid by the effect of upstream and downstream features along a flow path (such as dams and waterfalls) and influences arising from bioclimatic processes and human activities in the upstream catchment area. In the ‘spatially continuous longitudinal and lateral mosaics’ (Fausch et al. 2002, p. 3) created by these multi-scale processes, different elements that are critical for various fish life history stages are often separated in space and time (Kareiva and Wennergren 1995).

The importance of understanding processes and their interactions from multi-scale perspectives has been demonstrated and repeatedly emphasised (Ward 1989, Schlosser 1991, Jackson et al. 2001, Fausch et al. 2002, Durance et al. 2006, Lowe et al. 2006). Researchers have been urged to apply landscape ecology principles and insights to riverine environments (Fausch et al. 2002, Wiens 2002). These recommendations are more than academic, Fausch et al. (2002) noted that ecologists tend to work at small spatial and temporal scales, and have been generally ineffective at providing managers with information and tools at the landscape scales relevant to freshwater conservation. Durance et al. (2006, p. 1145) further criticise research for failing management by neither indicating ‘where in the scale hierarchy problems might arise nor the most effective scales for management response’.

Characterising the multidimensional environmental space of freshwater biota at multiple scales requires a broad range of predictors. A notable shortcoming of some studies that have attempted to predict the distributions of freshwater biota using variables at more than one spatial scale is that variables used have sometimes seemed more dictated by availability than functional relevance (see, e.g. Fransen et al. 2006). Austin (2002) has argued strongly for the use of theoretically informed, proximal variables that represent resource and direct gradients that influence species. Distal predictors (e.g. elevation) are less preferred because they only indirectly influence species distributions through their relationships with functionally relevant (proximal) predictors such as temperature. Recent examples show how effective purpose-built, ecologically relevant stream datasets can be for both landscape-scale strategic planning and finer-scale decision support for conservation management (e.g. Moilanen et al. 2008, 2011). Some base their analyses on environmental data from hierarchical stream networks that facilitate the tracing of upstream and downstream influences (e.g. Leathwick et al. 2010), whereas others have emphasised multi-scale landscape metrics calculated for individual sites at riparian, reach and subcatchment levels (e.g. Hopkins 2009). Both approaches provide the requisite multi-scale perspective that enables more integrative analyses for better management support. There has also been much interesting
work aimed at identifying meaningful and manageable attributes that can inform practical landscape management strategies (e.g. Clapcott et al. 2010, Hopkins and Whiles 2011).

This research was motivated by the need for context-appropriate quantitative tools to support conservation and management of biota in Australian freshwater environments. This article aims to (1) present our methods for constructing a fine-scale stream network GIS database for Victoria (south-eastern Australia), consisting of ecologically meaningful environmental attributes at multiple scales; (2) analyse patterns of distribution for a range of native and alien riverine fish species with diverse life-histories and habitat requirements and (3) demonstrate and briefly discuss applications of the resultant mapping of species distributions across stream networks.

Methods

Study area

The ∼127,500 km² study area comprises catchments of the inland-flowing river systems of Victoria (Figure 1). Major systems include the Upper Murray, Mitta Mitta, Kiewa, Ovens, Broken, Goulburn, Campaspe, Loddon, Avoca and Wimmera Rivers. These rivers flow into the Murray River, except for the Avoca and Wimmera, which terminate in

Figure 1. Geographic distribution of fish sample sites used in the analysis (open circles). Inset shows the location of the study area relative to the state of Victoria and Australia. The major northward-draining river systems in the study area are annotated. The Murray River collects the waters of the first eight river systems in turn as it flows westward. Major on- and off-stream storages associated with the major river systems are indicated in purple.
ephemeral wetlands. The landscapes and climates of the major catchments are detailed in Appendix 1 in Supplementary Material – in overview, headwaters arise in forested, mountainous or hilly areas and drain northward through increasingly agricultural land. Several on- and off-stream storages impact flows (Figure 1).

Fish data
Fish occurrence data (Figure 1) were obtained from the Victorian Department of Sustainability and Environment (DSE) (Aquatic Fauna Database, supplied 1 June 2007), the Murray-Darling Basin Commission (SRA Asset SRA524, supplied 4 May 2007) and Koster et al. (2006). All records came from sampling conducted between 1980 and 2006 (inclusive). After removing taxonomically uncertain records, pre-processing to remove conspicuous errors, and exhaustive manual cross-checking of sites against auxiliary information and spatial datasets to validate positional accuracy, 1906 sites were selected for the analysis. Presence–absence data were extracted for 17 fish species (11 native and non-diadromous, 6 alien) that occurred in the dataset with a capture frequency $\geq 3\%$ (Table 1).

| Code     | Species name (common name)                                                  | Prevalence |
|----------|-----------------------------------------------------------------------------|------------|
| GADBIS   | *Gadopsis bispinosus* Sanger 1984 (Two-spined blackfish)                    | 334/1906 = 0.175 |
| GADMAR   | *Gadopsis marmoratus* Richardson 1848 (River blackfish)                     | 243/1906 = 0.128 |
| GALFUS   | *Galaxias fuscus* Mack 1936 (Barred galaxias)                               | 73/1906 = 0.038 |
| GALOLI   | *Galaxias olidus* Günther 1842 (Mountain galaxias)                          | 350/1906 = 0.184 |
| MACMAC   | *Maccullochella macquariensis* Cuvier 1829 (Trout cod)                      | 62/1906 = 0.033 |
| MACPEL   | *Maccullochella peelii peeli* Mitchell 1838 (Murray cod)                    | 93/1906 = 0.049 |
| MACMAB   | *Macquaria ambigua* Richardson 1845 (Golden perch)                          | 137/1906 = 0.072 |
| MACAUS   | *Macquaria australasica* Cuvier 1830 (Macquarie perch)                      | 82/1906 = 0.043 |
| NANAUS   | *Nannoperca australis* Günther 1861 (Southern pygmy perch)                  | 98/1906 = 0.051 |
| PHIGRA   | *Philypnodon grandiceps* Krefft 1864 (Flathead gudgeon)                     | 126/1906 = 0.066 |
| RETSEM   | *Retropinna semoni* Weber 1895 (Australian smelt)                           | 204/1906 = 0.107 |
| CARAUR   | *Carassius auratus* Linnaeus 1758 (Goldfish)                                | 207/1906 = 0.109 |
| CYPCAR   | *Cyprinus carpio* Linnaeus 1758 (Common carp)                               | 338/1906 = 0.177 |
| GAMHOL   | *Gambusia holbrooki* Girard 1859 (Eastern gambusia)                          | 151/1906 = 0.079 |
| ONCMYK   | *Onchorynchus mykiss* Walbaum 1792 (Rainbow trout)                          | 358/1906 = 0.188 |
| PERFLU   | *Perca fluviatilis* Linnaeus 1758 (Redfin perch)                            | 507/1906 = 0.266 |
| SALTRU   | *Salmo trutta* Linnaeus 1758 (Brown trout)                                 | 746/1906 = 0.391 |

Notes: L, listed as threatened under the Victorian Flora and Fauna Guarantee (FFG) Act 1998; EN, listed as endangered under the Commonwealth Environment Protection and Biodiversity Conservation (EPBC) Act 1999; VU, listed as vulnerable under Commonwealth EPBC Act 1999. Native species are shown in boldface.
Four of the species are considered to be threatened at the State and/or Commonwealth level (Table 1).

**Environmental data**

In terrestrial environments, the unit for spatially explicit landscape planning/management is typically a grid cell. As this is inappropriate for riverine systems, analysts tend to employ grid or polygon-based representations of subcatchments/watersheds (e.g. Linke et al. 2008). These are convenient because species records, when mapped, often do not coincide exactly with mapped linework for rivers. While this might suffice for broad-scale visualisation and planning tasks (e.g. at scales of 10s–100s km²), it is generally too coarse for the scales at which on-ground managers operate. When target entities are riverine species/communities, stream-based representations such as the US National Hydrography Dataset (http://nhd.usgs.gov), Australia’s Geofabric Surface Network (www.bom.gov.au/water/geofabric/about.shtml), and the stream network described in this article are more pertinent.

The backbone of our digital stream network database is an ordered, link-node representation of Victoria’s streams. As detailed in Appendix 2A in Supplementary Material, we constructed a 20 × 20 m cell-size digital elevation model (DEM) and from this produced a stream network with flow directionality, connectivity and associated watersheds for each of the 238,474 links in the study area. To characterise the riverine environments, estimates of physiographic, bioclimatic, edaphic, land-cover- and human-disturbance-related variables considered to have ecological relevance for freshwater biota, were computed for every stream link at one or more hierarchically nested spatial scales (Appendix 2B in Supplementary Material). The three scales were (1) the riparian zone with a width of 50 m on either side of a link; (2) the immediate watershed of a link and (3) the entire upstream contributing catchment area (UCA) associated with a link (see Figure 2).

Inputs for environmental variables were generated from best-available data or derived from the most detailed and up-to-date sources for Victoria at the time of construction. Physiographic variables were calculated directly from the DEM or indirectly, by calculating terrain attributes including topographic wetness index (Moore et al. 1993) and multi-resolution valley bottom flatness index (Gallant and Dowling 2003). Bioclimatic variables were derived from interpolated surfaces estimated using the software package ANUCLIM 5.1 (Houlder et al. 2000), which uses thin-plate smoothing splines fitted to long-term meteorological station data. Edaphic variables were calculated from modelled solum depth and plant available water holding capacity extracted from the Soil Hydrological Properties of Australia spatial dataset (Western and McKenzie 2006). Land-cover variables were derived from the Modelled Native Vegetation Extent spatial dataset (NVE2007, DSE). For disturbance-related variables, an input raster of road density (km/km²) was created using ArcGIS tools and road features captured at a 1:25,000 scale (VicMap Transport, DSE).

Georeferenced data representing in-stream structures/features of potential influence on fish, such as rapids, waterfalls, dam walls, gauging stations and fords were obtained from VicMap Hydro and Transport and Thiess Services Pty Ltd. These features were spatially joined to the stream network, explicitly flagged in each affected link and used in computing other variables (see below).

Variable estimates for each link were computed at the riparian, watershed and UCA scales using a suite of custom ArcINFO scripts. Variables at watershed-scale were quantified by overlaying watershed boundaries on the various environmental datasets.
Figure 2. Example showing nested spatial scales at which environmental variables for individual links were computed. Flow direction is to the north. For the highlighted link, the hatched area indicates the link’s riparian zone, the area enclosed by the dotted line denotes the link’s immediate watershed (the upslope area that contributes flow to the link) and the solid shaded area represents the entire upstream contributing catchment area (UCA) for the highlighted link. The UCA encompasses the ‘immediate watershed’ areas for all contributing links upstream of the highlighted link.

Riparian-scale variables were computed in a similar manner using buffer boundaries generated along each stream link. UCA-scale variables were computed from watershed- or riparian-scale variables using network accumulation algorithms. The general modelling approach for these operations followed that of Wilkinson et al. (2004). Finally, the directionality of the stream network was exploited to conduct ‘traces’ in upstream–downstream directions along the flow path of each link, computing a range of variables of potential ecological relevance using purpose-written tracing scripts. For example, estimates of the average slope encountered along a link’s upstream flow path or the maximum slope encountered along a link’s downstream flow path (US_AVGSLOPE and DS_MAXSLOPE, respectively, in Table 2). These geoprocessing operations produced the fluvial equivalent of neighbourhood metrics for focal cells in terrestrial settings. The current number of available environmental predictors is 96 (Appendix 2B in Supplementary Material), but only a subset was used for modelling (Table 2).

The spatially corrected fish sample sites were spatially joined to the network (tolerance of 40 m) to extract estimates of predictors from corresponding stream links. We reviewed the literature (Jackson and Williams 1980; Cadwallader and Backhouse 1983; McDowall 1996, 2006, Morris et al. 2001, Allen et al. 2002) for ecological insights on proximal and distal predictors to guide our initial selection for the 17 target species (Williams et al. 2012). We concentrated on predictors that are temporally stable over the time period.
Table 2. Descriptions of candidate environmental and aspatial variables used in the analysis.

| Variable | Description | Mean, range (missing) |
|----------|-------------|-----------------------|
| **Physiographic predictors** | | |
| LINK_UCA_HA | Link UCA area (ha) | 55608, 0.6–911509 (125) |
| SLOPE_PERC | Link slope calculated as rise/run * 100 (%) | 2.59, 0.00–47.44 |
| DS_MAXLSOPE | Maximum slope encountered along link downstream flow path (%) | 9.74, 0.00–71.65 (106) |
| US_AVGSLOPE | Average slope encountered along link upstream flow path (%) | 9.45, 0.00–46.20 (125) |
| UC_TWI | Mean topographic wetness index (TWI) value in link UCA | 9.7, 7.6–19.1 (125) |
| FLOODWIDTH | Link floodplain width (km) | 0.22, 0.00–22.53 |
| **Bioclimatic predictors** | | |
| UC_MEAN_PPT | Mean total annual precipitation in link UCA (mm/yr) | 1142, 338–2392 (125) |
| UC_CV_PPT | Mean coefficient of variation of mean annual precipitation in link UCA | 28.6, 14.0–37.0 (125) |
| UC_MEAN_TEMP | Mean annual temperature in link UCA | 11.0, 5.1–15.6 (125) |
| UC_CV_TEMP | Mean coefficient of variation in mean annual temperature in link UCA | 1.63, 1.42–1.92 (125) |
| WETQT_TEMP | Mean temperature of wettest quarter (any 13 consecutive weeks) in link watershed | 7.0, 0.0–12.2 |
| WARMQT_TEMP | Mean temperature of warmest quarter (any 13 consecutive weeks) in link watershed | 18.8, 11.2–22.4 |
| **Edaphic predictors** | | |
| UC_SOLDEPTH | Mean solum depth in link UCA (m) | 1.09, 0.50–1.50 (125) |
| **Landcover predictors** | | |
| NONATVEG | Mean proportion of link watershed without any native vegetation cover | 0.29, 0.0–1.00 |
| UC_NONATVEG | Mean proportion of link UCA without any native vegetation cover | 0.21, 0.00–0.96 (125) |
| RIP_NONATVEG | Mean proportion of riparian zone in link watershed without any native vegetation cover | 0.16, 0.00–1.00 |
| TRECOV | Mean proportion of link watershed covered by any type/combination of tree cover | 0.59, 0.00–1.00 |
| RIP_TRECOV | Mean proportion of riparian zone in link watershed covered by any type/combination of tree cover | 0.70, 0.00–1.00 |
| UC_TRECOV | Mean proportion of link UCA covered by any type/combination of tree cover | | |
| UC_ROAD_DENS | Road density in link UCA (km/km²) | 1.40, 0.00–7.56 (125) |
| **Aspatial predictors** | | |
| SURV_METH | Fish survey methods in five classes (numbers in parentheses denote total number of sites associated with each survey method): electrofishing (970), netting (409), traps (226), mixed (41) and other (260) | Not applicable |

Notes: See Appendix 2B in Supplementary Material for the full set of available environmental variables. Mean and range values are for the sites used in the analysis, not the entire stream network in the study area. A number of sites had missing values for certain predictors which could not be estimated because either the upstream contributing catchment area or the downstream flow path exceeded the extent of the stream network. The total number of such sites in relation to each predictor is indicated in parentheses.
that the fish data were collected. Predictors used for modelling had pairwise Pearson correlations < 0.85.

**Statistical modelling: boosted regression trees**

We selected boosted regression trees (BRTs) to analyse the relationships between records of whether a species was caught or not, and the environmental attributes of the associated stream link. Details of the method are published (Elith *et al.* 2008, Hastie *et al.* 2009) so here we give a brief overview (see Appendix 3 in Supplementary Material for elaboration) and details of our modelling choices. BRTs are a combination of two algorithms: regression trees (a type of decision tree) and boosting (a forward, stagewise model-fitting procedure). The final BRT model can be understood as an additive regression model (Friedman *et al.* 2000) in which individual terms are trees (sometimes thousands of them) which have been fitted sequentially until an optimal level of complexity is achieved. BRTs are increasingly used in ecology, with applications spanning SDMs and various other regression-style analyses (examples and references in Appendix 3 in Supplementary Material). Relative importance of variables, partial dependence plots (showing the marginal effect of a variable on the response after accounting for the average effects of all other variables in the model) and prediction can all be estimated for BRTs (Elith *et al.* 2008 and Appendix 3 in Supplementary Material).

All analyses were carried out in R (R Foundation for Statistical Computing 2008) using the ‘gbm’ package (v1.6–3, Ridgeway 2007) plus additional code written by Elith *et al.* (2008). BRT models were fitted using a learning rate of 0.01 and a tree complexity of three nodes, allowing up to three-way interactions among predictors. We used these settings as a starting point and checked the number of trees in resultant models, adjusting the learning rate downwards where necessary, to ensure final models with around 1000 trees or more. This mitigates against unwanted variation between runs of the stochastic modelling process (Elith *et al.* 2008, Appendix 1). To control for overfitting, we used 10-fold cross-validation (CV) for model development and automatic selection of the optimal number of trees. For each species model, a final set of predictors were selected from the full candidate set using model simplification code that sequentially dropped the least important predictors, until the model was optimised for minimal average CV error in the 10-fold CV process (Elith *et al.* 2008).

In addition to abiotic predictors, we considered biotic interactions. Including the presence of other species as predictors might be informative regarding patterns of co-occurrence or avoidance for native species. We reviewed literature on geographic distribution and potential species interactions to identify the list of species thought to influence each of the 11 native species (Appendix 4A in Supplementary Material).

If predictive performance is of interest, model evaluation should target independent sites not used in model training, and a common method is to use sites set aside during CV (Leathwick *et al.* 2008a). We did this using CV folds stratified by prevalence, utilising the model building CV to provide measures of predictive performance (as detailed in Elith *et al.* 2008). This gives an indication of predictive performance in the same general set of rivers, which is primarily what we are interested in here. However, the hierarchical structure of river systems means that some sites within a given CV fold will be along the same river, and therefore this evaluation approach will be more optimistic than one that tests the model on rivers outside the training sample. To assess model performance at spatially independent sites (‘transferability’), we assembled data for 2492 sites across southern Victoria. This is a difficult test, because the south-draining river systems traverse different landscapes
and have substantially different disturbance histories compared with the north-draining ones used for model training. We could evaluate 11 species models using these data; the remaining species were too rare in these southern rivers for an informative result.

We used two model evaluation metrics – the percentage deviance explained and the area under the receiver operating characteristic curve (AUC). Deviance explained measures the goodness-of-fit between predicted and raw values. We express it as a percentage of the null deviance (the deviance of a model containing no terms and having a fitted value for all observations equal to the mean probability in the observations) for each species. AUC measures a model’s ability to discriminate between sites where the species is present and where it is absent. The AUC value is equivalent to the probability that a randomly selected presence record will have a higher fitted probability value than a randomly chosen absence record. AUC ranges from 0 to 1–1 indicating perfect discrimination while 0.5 implies predictive ability that is no better than a random guess. We estimated deviance explained and AUC for cross-validated models, and AUC for prediction to the geographically distinct southern rivers.

Results

Since observations are the presence or absence of the species conditional on survey method, and repeat visit data were unavailable for estimating detection probabilities, we report the results as probability of observation or catch. While the survey method variable adjusts for the variation in observation across methods, it does not allow absolute statements about probabilities of occurrence. Model summaries and performance metrics for each species are presented in Table 3. The number of variables retained in each final model ranged from nearly all variables retained in the case of *Maccullochella macquariensis* to only five for *Carassius auratus* (Table 3). Mean cross-validated percentage deviance explained by models was slightly higher for native species (42.8%) than for alien species (35.3%). Native species models performed slightly better at discriminating between observed presences and absences than models for alien species (mean cross-validated AUC of 0.928 versus 0.888).

From each species model, we extracted the top ten predictors based on their average influence (Appendix 5 in Supplementary Material). Functions fitted by BRT models for all species are presented in Appendix 6 in Supplementary Material. Top predictors for each species span a range of spatial scales from the riparian zone to the link watershed and UCA (Appendix 5 in Supplementary Material). For 6 of the 17 species, catch response was strongly dependant on the survey technique (SURV_METH is ranked among the top three predictors for these species, Appendix 5 in Supplementary Material).

Including the presence of other species, turned out to be influential for *Gadopsis bispinosus, Galaxias fuscus* and *Macquaria ambigua* (Appendix 5 in Supplementary Material). The following results are based on relationships evident in fitted functions for each species (see Appendix 6 in Supplementary Material). The presence of *Cyprinus carpio* was the strongest correlate of distribution for *M. ambiguа*. Observations of *Macullochella peelli peelli* were also positively correlated with the presence of *C. carpio*, but the marginal effect was very small. *G. fuscus* catch was negatively correlated with the presence of *Salmo trutta* but the converse was true with respect to *G. bispinosus*.

To illustrate the relationships modelled with these data, we present fitted functions for the top four predictors in models for *G. bispinosus, G. fuscus* and *C. carpio* (Figure 3). Here, we describe the main trends, particularly where data are more dense (see ticks at the top of each pane, Figure 3). The partial plots for *G. bispinosus* (Figure 3a) indicate that it is caught most frequently in links where the maximum slope along downstream flow paths.
## Table 3. Comparison of the predictive performance of BRT models.

| Species         | No. of trees | No. of variables | Null deviance | CV predictive deviance (s.e.) | CV deviance explained (%) | AUC (s.e.)  |
|-----------------|--------------|------------------|---------------|--------------------------------|----------------------------|-------------|
| G. bispinosus   | 1300         | 19               | 0.928         | 0.457 (0.023)                  | 50.7                       | 0.944 (0.006) |
| G. marmoratus   | 2450         | 19               | 0.763         | 0.444 (0.020)                  | 41.8                       | 0.917 (0.011) |
| G. fuscus       | 1650         | 20               | 0.325         | 0.135 (0.006)                  | 58.5                       | 0.976 (0.005) |
| G. olidus       | 2950         | 19               | 0.954         | 0.729 (0.020)                  | 23.6                       | 0.833 (0.010) |
| M. macquartienis| 950          | 21               | 0.287         | 0.153 (0.016)                  | 46.8                       | 0.942 (0.015) |
| M. peelli peelli| 1000         | 20               | 0.390         | 0.217 (0.018)                  | 44.4                       | 0.957 (0.009) |
| M. ambigua      | 1900         | 21               | 0.517         | 0.240 (0.012)                  | 54.3                       | 0.968 (0.003) |
| M. australasica | 1600         | 14               | 0.355         | 0.194 (0.014)                  | 45.4                       | 0.945 (0.010) |
| N. australis    | 1700         | 9                | 0.405         | 0.258 (0.016)                  | 36.3                       | 0.919 (0.014) |
| P. grandiceps   | 950          | 17               | 0.487         | 0.303 (0.014)                  | 37.7                       | 0.929 (0.008) |
| R. semoni       | 1700         | 15               | 0.678         | 0.469 (0.019)                  | 30.8                       | 0.882 (0.015) |
| C. auratus      | 1350         | 5                | 0.687         | 0.502 (0.011)                  | 27.0                       | 0.866 (0.007) |
| C. carpio       | 1450         | 7                | 0.935         | 0.510 (0.019)                  | 45.4                       | 0.928 (0.006) |
| G. holbrooki    | 1150         | 9                | 0.554         | 0.371 (0.012)                  | 33.0                       | 0.891 (0.010) |
| O. mykiss       | 1750         | 20               | 0.966         | 0.622 (0.028)                  | 35.7                       | 0.888 (0.009) |
| P. fluviatilis  | 1450         | 20               | 1.158         | 0.768 (0.018)                  | 33.7                       | 0.875 (0.008) |
| S. trutta       | 2550         | 11               | 1.339         | 0.847 (0.023)                  | 36.7                       | 0.881 (0.007) |

Range (Mean) 950–2950 (1638) 5–21 (15.65) 0.287–1.34 (0.690) 0.135–0.847 (0.425) 23.6–58.5 (40.1) 0.833–0.976 (0.914)

Note: Native species shown in boldface. Table entries show for each species model, the final number of trees fitted, final number of predictor variables retained, null deviance, and cross-validated estimates and standard errors (s.e) of mean predictive deviance, mean percentage of the total deviance explained and mean AUC. The learning rate used for all species was 0.01, except for M. ambigua, R. semoni, C. auratus and G. holbrooki where the learning rate used was 0.005.
Figure 3. Partial dependence plots for the four most influential variables in the BRT models for (a) *G. bispinosus*, (b) *G. fuscus* and (c) *C. carpio*. For explanations of variables and their units, see Table 2. Values in parentheses refer to the relative percentage contribution of the variable in the species model. A common scale is used for the y-axis of plots for each respective species. Ticks across the top edge of each plot indicate the distribution of values for each continuous variable in deciles.
is >30%. It also tends to be caught in links closer to headwater areas where the proportion of tree cover in the UCA is >0.6 and where the mean UCA precipitation is >1000 mm and watershed temperature in the warmest quarter is >16.5°C. G. fuscus is most likely to be caught in links where the UCA is nearly completely tree covered and where watershed temperature in the wettest quarter is <≈4.5°C and in the warmest quarter is <≈16.5°C (Figure 3b). C. carpio is detected most frequently using electrofishing, netting and mixed methods and in links with where the UCA is large, link slope is <0.1% and watershed temperature in the warmest quarter is >≈20°C (Figure 3c).

Figure 4 illustrates mapped model predictions for the introduced S. trutta and two native blackfish, Gadopsis marmoratus and G. bispinosus, within the catchment of the upper Goulburn River. Until relatively recently, G. marmoratus and G. bispinosus were considered a single species (Lintermans 2000). However, when the relationships in environmental space were modelled in our analysis and mapped, a quite distinct sorting of G. marmoratus and G. bispinosus in geographic space becomes apparent (Figure 4b and c). In river systems where both species have a high predicted probability of observation G. marmoratus is replaced by G. bispinosus in the upper tributaries. S. trutta which is thought to have adverse effects on both species via competition and predation has a high predicted probability of catch over considerable lengths of streams across the illustrated region (Figure 4a). In particular, there is a high degree of overlap with the predicted distribution of G. bispinosus.

Species prevalence in the south-draining rivers differed from that in the north (Tables 1 and 4), with the native species being considerably more rare in the southern catchments, and the introduced species more common for all but one species. When models developed with data from north-draining systems were used to predict to sites in south-draining rivers, model performance ranged from little better than a random guess in the case of G. marmoratus to useful discriminatory ability with respect to Philypnodon grandiceps, C. carpio and Gambusia holbrooki (Table 4). Models for alien species were better at discriminating between observed presences and absences than native species models (mean AUC of 0.724 versus 0.633), and the decrease in AUC from that estimated in the cross-validated data ranged from 0.14 to 0.41 for native species and 0.12 to 0.21 for introduced species.

Discussion

Stream-link-based spatial infrastructure

Using commercially available and custom-written GIS geoprocessing tools, we created a fine-scale stream network GIS database and populated it with a suite of physiographic, bioclimatic, edaphic and land-cover variables estimated at multiple scales. Calculating multi-scale metrics for individual sites is an alternative means of providing the desired multi-scale perspective for analysis. However, we opted to develop a hierarchical network-based system because once built, it can be used with any set of input sites, can be extended to include additional attributes, gives greater flexibility for analysing upstream–downstream effects on a site, and for assessing effects of management and conservation priorities in a framework consistent with the functional connectivity pattern of river systems (see, e.g. Moilanen et al. 2011).

This GIS database constitutes primary spatial infrastructure that can support the type of multi-scale research and management of riverine ecosystems advocated by Fausch et al. (2002), Durance et al. (2006) and Lowe et al. (2006). It has already been used to develop freshwater SDMs for a state-wide spatial biodiversity prioritisation map (NaturePrint v2.0, DSE 2012) and is being used to develop a quantitative classification of riverine
Figure 4. Maps of predicted probability of catch for (a) *S. trutta*, (b) *G. marmoratus* and (c) *G. bispinosus* in the upper Goulburn region.
Table 4. Prevalence of the 11 fish species across 2492 sites in the geographically independent south-draining catchments and AUC scores when models developed using sites in north-draining catchments ($N = 1906$) are used to predict to sites in south-draining river systems.

| Species       | Prevalence | AUC       |
|---------------|------------|-----------|
| *G. marmoratus* | 519/2492 = 0.208 | 0.507     |
| *G. olidus*   | 329/2492 = 0.132 | 0.623     |
| *N. australis*| 418/2492 = 0.168 | 0.640     |
| *P. grandiceps* | 308/2492 = 0.124 | 0.785     |
| *R. semoni*   | 484/2492 = 0.194 | 0.606     |
| *C. auratus*  | 175/2492 = 0.070 | 0.718     |
| *C. carpio*   | 138/2492 = 0.055 | 0.768     |
| *G. holbrooki*| 208/2492 = 0.083 | 0.775     |
| *O. mykiss*   | 75/2492 = 0.030  | 0.678     |
| *P. fluviatilis* | 258/2492 = 0.104 | 0.710     |
| *S. trutta*   | 639/2492 = 0.256 | 0.693     |
| Range (mean)  | 0.030–0.256 (0.130) | 0.507–0.785 (0.682) |

Note: Native species shown in boldface.

habitat types (Bill O’Connor, personal communication, DSE). While the GIS database includes the main biophysical attributes likely to influence biodiversity, we recognise that to be of maximal utility for policy and management, it ought to include variables that are amenable to management (Hopkins and Whiles 2011). Currently available predictors that are amenable to management relate mainly to land-cover (e.g. proportion of riparian tree cover) and in-stream structures (e.g. number of dam walls along a link’s upstream/downstream flow path) (Appendix 2B in Supplementary Material). Predictors under development that will be management-relevant include a range of hydrological indices to represent various aspects of the flow regime and presence–absence of in-stream coarse woody debris.

**Evaluating the SDMs**

For modelling fish species, we chose candidate predictors based on their likely ecological relevance. Upstream catchment area (LINK_UCA_HA) featured among the top three predictors for 11 out of the 17 species and may be important in indicating longitudinal position of a site within a catchment. The variable sets that were identified as influential and the relationships fitted to them differed from species to species with no obvious dominant set of predictors explaining variation in the probability of catch. This may be a function of the modest environmental contrasts in the weathered landscapes of the study area. Influential predictors for each species spanned the full range of spatial scales and included terms relating to catch method, site location within a catchment, morphological, land-cover and bioclimatic attributes. Together, they describe the mesoscale habitat characteristics of each species. For some species, the variables and fitted relationships had a relatively straightforward interpretation with respect to published accounts of a species’ preferred habitat. For instance, this applied to *M. peelii peelii* and *C. carpio*, whose fitted relationships (middle order to lowland catchment position, very low gradient slopes and warm temperatures in the warmest quarter) matched well with published descriptions of habitat preferences derived from field studies (i.e. slow-flowing or still waters with warm water temperatures during spring–summer; Rowland 1988, Gilligan and Rayner 2007). For other species, the
variable set identified as influential was less obvious but suggested a range of ecologically plausible interpretations.

Negative interactions with alien fish species are often cited as threats contributing to native species’ decline in range and/or abundance. However, experimental studies allowing rigorous testing of these hypotheses are few and challenging to implement under natural conditions. Available evidence is mostly circumstantial; nevertheless, some interactions such as the adverse impacts of trout on galaxiids have been exhaustively documented (McDowall 2006). We included the presence of potential competitors/predators as additional predictors because we wanted to predict species distributions as accurately as possible. Our intent in including potential competitors was not to attempt to make definitive statements about the nature of biotic interactions between species; with these data alone, it is not possible to clearly attribute the relationship modelled between the target species and its competitor to direct biotic interaction. Some of the species models might include the competitor simply because it is an efficient way to represent environments important to the presence or absence of the target species. Furthermore, the competitor’s distribution might indicate site-scale conditions or microhabitat characteristics not represented in our predictor set. Rarely can biotic interactions be clearly identified from SDMs (but see Leathwick and Austin (2001) for an exception). Indeed, in our models, the responses modelled for competitors ranged from ecologically plausible to counter intuitive. The conventional expectation of a negative relationship between the presence of S. trutta on G. fuscus occurrence (Koehn and O’Connor 1990, McDowall 2006) was realised (Appendix 6 in Supplementary Material). The presence of potential predators/competitors had no detectable influence on the modelled catch of G. marmoratus, Galaxias olidus, M. macquariensis, M. peelii peelii, Macquaria australasica, Nannoperca australis and P. grandiceps. For two species apparently counter-intuitive results were obtained — that is, positive associations between: the introduced S. trutta and G. bispinosus, and between the introduced C. carpio and M. ambigua (Appendices 5 and 6 in Supplementary Material). Nevertheless, these models appear to predict well within the region in which the models were developed (on held-out, cross-validated data, ∼50% deviance explained and AUC ∼0.95, Table 3). Neither of these native species had sufficient data in the geographically independent dataset to test transferability.

The use of potential predators/competitors in these models is a possibly valid approach that requires further testing. This might include predicting to various river systems, comparing the results from models with and without the competitors, and testing the modelled results via additional surveys or expert evaluation (see Appendix 4B in Supplementary Material for further discussion).

As demonstrated, the modest environmental contrasts of the study area posed no particular obstacle to developing models with valuable predictive ability and discriminatory power in the training region. There is growing interest in assessing model transferability (e.g. Heikkinen et al. 2012; Wenger and Olden 2012) and this study’s predictions to the southern rivers adds an interesting (and rare) example evaluation of the ability of the models to predict to geographically independent rivers. For the 11 species tested, the models always had a lower ability to discriminate between catch and non-catch in these southern rivers (compare AUCs in Tables 3 and 4). This is understandable, as these rivers are in substantially different landscapes with different disturbance histories. The lower prevalence of native species and higher prevalence of introduced species in southern rivers are indicators of these different conditions. Most models retained some discriminatory ability in the southern region. In reality, if predictions were required for these southern regions, models would either be best fitted to state-wide datasets or perhaps specifically to the
south-draining rivers. Nevertheless, being able to evaluate predictions in rivers not used in model training is a useful test, and where data can be sensibly subdivided by river basins for model training and testing, this may be preferable for many applications. In our case, this was not feasible because over half the species analysed were patchily distributed across the river basins.

**Applications of mapped SDMs**

A particular advantage of a fine-scale, stream-link-based network over larger-scale, areal-based representations is that predictions can be made at scales that can be feasibly validated. Mapping quantitative models to specific spatial contexts, enables biologists to visually compare their mental spatial models of fish occurrence with those developed mathematically (McCleary and Hassan 2008). This allows expert evaluation of the models, identification of potential errors and enhancements, all of which are important to iterative modelling processes. Predictions for unsampled streams provides information that can assist ecological understanding, inquiry and management. For instance, mapped predictions revealed the spatial sorting of the closely related *G. marmoratus* and *G. bispinosus* (Figure 4b and c). Predictions can help identify sites of high potential occurrence of rare species, suitable sites for establishing new populations of valued species and sites at risk of invasive species proliferation. Species-by-species predictions for a region of interest provides a robust method of deriving species assemblage profiles and provides a more informative context for interpreting monitoring results (Magness *et al.* 2008). Using the full capability of a GIS, mapped predictions can be summarised at any scale of management interest and used in conjunction with conservation planning tools for tasks such as quantitatively assessing the biodiversity conservation value of streams (see Moilanen *et al.* 2008, Leathwick *et al.* 2010, DSE 2012), and prioritising conservation and/or rehabilitative efforts (see Moilanen *et al.* 2011). Both the GIS stream network database and SDMs have immediate applications, but they also provide a firm foundation for the development of more sophisticated tools for addressing challenges in the management and conservation of Australian freshwater environments.

**Acknowledgements**

YEC was funded by ARC Linkage LP0667891 and LP110100304 and JE by ARC DP0772671 and FT0991640. We thank Ron De Rose, Matt White, John Leathwick, Phil Scanlon, John Gallant and Nada Filipovic for their advice and assistance in constructing the GIS stream network database. We also thank Kelly Hayles, Greg Long and Wayne Koster for access to the fish data and Will Morris for data preparation. John Leathwick, Simon Nicol, Peter Vesk, Ivars Reinfelds, Shawn Laffan and two anonymous reviewers provided comments that improved the manuscript.

**References**

Allen, G.R., Midgeley, S.H., and Allen, M., 2002. *Field guide to the freshwater fishes of Australia*. Perth: West Australian Museum.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157, 101–118.

Cadwallader, P.L. and Backhouse, G.N., 1983. *A guide to the freshwater fish of Victoria*. Melbourne: Department of Conservation.

CES, Commissioner for Environmental Sustainability, 2008. *State of the environment Victoria 2008*. Melbourne: Commissioner for Environmental Sustainability.
Clapcott, J.E., et al., 2010. Exploring the response of functional indicators of stream health to land-use gradients. *Freshwater Biology*, 55 (10), 2181–2199.

DSE, Department of Sustainability and Environment, 2012. *NaturePrint: freshwater species distribution models* [online]. Available from: http://www.dse.vic.gov.au/conservation-and-environment/biodiversity/natureprint/natureprint-products

Durance, I., Lepichon, C., and Ormerod, S.J., 2006. Recognizing the importance of scale in the ecology and management of riverine fish. *River Research and Applications*, 22 (10), 1143–1152.

Elith, J., Leathwick, J.R., and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802–813.

Engler, R., Guisan, A., and Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41, 263–274.

Fausch, K.D., et al., 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *BioScience*, 52 (6), 483–498.

Fransen, B.R., et al., 2006. A logistic regression model for predicting the upstream extent of fish occurrence based on geographical information systems data. *North American Journal of Fisheries Management*, 26, 960–975.

Friedman, J.H., Hastie, T., and Tibshirani, R., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802–813.

Gallant, J.C. and Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research*, 39 (12), 1347–1360.

Gilligan, D. and Rayner, T., 2007. The distribution, spread, ecological impacts and potential control of carp in the upper Murray river. Cronulla: NSW Department of Primary Industries.

Hartley, S., Harris, R., and Lester, P.J., 2006. Quantifying uncertainty in the potential distribution of an invasive ant species: climate and the Argentine ant. *Ecology Letters*, 9 (9), 1068–1079.

Hastie, T., Tibshirani, R., and Friedman, J., 2009. *The elements of statistical learning*. 2nd ed. New York: Springer.

Hauser, C.E. and McCarthy, M.A., 2009. Streamlining ‘search and destroy’: cost effective surveillance for invasive species management. *Ecology Letters*, 12 (7), 683–692.

Heikkinen, R.K., Marmion, M., and Luoto, M., 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography*, 35, 276–288.

Hopkins, R.L., 2009. Use of landscape pattern metrics and multiscale data in aquatic species distribution models: as case study of a freshwater mussel. *Landscape Ecology*, 24 (7), 943–955.

Hopkins, R.L. and Whiles, M.R., 2011. The importance of land use/land cover data in fish and mussel conservation planning. *Annales de Limnologie – International Journal of Limnology*, 47 (3), 199–209.

Houlder, D.J., et al., 2000. *ANUCLIM user guide version 5.1*. Centre for Resource and Environmental Studies. Canberra: Australian National University.

Jackson, D.A., Peres-Neto, P.R., and Olden, J.D., 2001. What controls who is where in freshwater fish communities – the roles of biotic, abiotic, and spatial factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 58, 157–170.

Jackson, P.D. and Williams, W.D., 1980. Effects of brown trout, *Salmo trutta* L., on the distribution of some native fishes in the three areas of southern Victoria. *Australian Journal of Marine and Freshwater Research*, 31, 61–67.

Kareiva, P. and Wennergren, U., 1995. Connecting landscape patterns to ecosystem and population processes. *Nature*, 373, 299–302.

Koehn, J.D. and O’Connor, W.G., 1990. *Biological information for management of native freshwater fish in Victoria*. Melbourne: Department of Conservation and Environment.

Koster, W.M., et al., 2006. *Surveys of fish communities in the lower Goulburn river*. Heidelberg, VIC: Department of Sustainability and Environment.

Leathwick, J.R., et al., 2008a. Dispersal, disturbance and the contrasting biogeographies of New Zealand’s diadromous and non-diadromous fish species. *Journal of Biogeography*, 35, 1481–1497.

Leathwick, J., et al., 2008b. Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters*, 1, 91–102.

Leathwick, J.R., et al., 2010. Complementarity-based conservation prioritization using a community classification, and its application to riverine ecosystems. *Biological Conservation*, 143, 984–991.
Leathwick, J.R. and Austin, M.P., 2001. Competitive interactions between tree species in New Zealand’s old-growth indigenous forests. *Ecology*, 82 (9), 2560–2573.

Linke, S., Norris, R.H., and Pressey, R.L., 2008. Irreplaceability of river networks: towards catchment-based conservation planning. *Journal of Applied Ecology*, 45, 1486–1495.

Linternans, M., 2000. The status of fish in the Australian capital territory: a review of current knowledge and management requirements. Canberra: Environment ACT.

Lowe, W.H., Likens, G.E., and Power, M.E., 2006. Linking scales in stream ecology. *BioScience*, 56 (7), 591–597.

Magnness, D.R., Huettmann, F., and Morton, J.M., 2008. Using random forests to provide predicted species distribution maps as a metric for ecological inventory and monitoring programs. *Studies in Computational Intelligence*, 122, 209–229.

Martínez-Meyer, E., *et al.*, 2006. Ecological niche modelling and prioritizing areas for species reintroductions. *Oryx*, 40 (4), 411–418.

McCleary, R.J. and Hassan, M.A., 2008. Predictive modeling and spatial mapping of fish distributions in small streams of the Canadian Rocky Mountain foothills. *Canadian Journal of Fisheries and Aquatic Sciences*, 65, 319–333.

McDowall, R.M., 1996. *Freshwater fishes of south-eastern Australia*. Sydney: Reed Pty Ltd.

McDowall, R.M., 2006. Crying wolf, crying foul, or crying shame: alien salmonids and a biodiversity crisis in the southern cool-temperate galaxioid fishes? *Reviews in Fish Biology and Fisheries*, 16 (3–4), 233–422.

Moilanen, A., Leathwick, J., and Elith, J., 2008. A method for spatial freshwater conservation prioritization. *Freshwater Biology*, 53 (3), 577–592.

Moilanen, A., Leathwick, J.R., and Quinn, J.M., 2011. Spatial prioritization of conservation management. *Conservation Letters*, 4 (5), 383–393.

Moore, I.D., *et al.*, 1993. Soil attribute predictions using terrain analysis. *Soil Science Society of America Journal*, 57, 443–452.

Morris, S.A., *et al.*, 2001. Threatened and potentially threatened freshwater fishes of coastal New South Wales and the Murray-Darling Basin. Sydney: NSW Fisheries.

Olden, J.D. and Jackson, D.A., 2002. A comparison of statistical approaches for modelling fish species distribution. *Freshwater Biology*, 47 (10), 1976–1995.

Ridgeway, G., 2007. *Generalized boosted models: a guide to the gbm package* [online]. Available from: http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf

Rowland, S., 1988. Murray cod. Agfact F3.2.4. Sydney: NSW Agriculture and Fisheries.

Schlosser, I.J., 1991. Stream fish ecology: a landscape perspective. *BioScience*, 41 (10), 704–712.

Steel, E.A., *et al.*, 2004. Landscape models to understand steelhead (*Oncorhynchus mykiss*) distribution and help prioritize barrier removals in the Willamette basin, Oregon, USA. *Canadian Journal of Fisheries and Aquatic Sciences*, 61, 999–1011.

Thuiller, W., *et al.*, 2008. Predicting global change impacts on plant species’ distributions: future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, 9 (3–4), 137–152.

Ward, J.V., 1989. The four-dimensional nature of lotic ecosystems. *Journal of the North American Benthological Society*, 8 (1), 2–8.

Wenger, S.J. and Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3 (2), 260–267.

Western, A. and McKenzie, N., 2006. Soil hydrological properties of Australia. Melbourne: CRC for Catchment Hydrology.

Wiens, J.A., 2002. Riverine landscapes: taking landscape ecology into the water. *Freshwater Biology*, 47 (4), 501–515.

Wilkinson, S., *et al.*, 2004. *SedNet user guide*. Canberra: CSIRO Land and Water.

Williams, K.J., *et al.*, 2012. Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, doi:10.1080/13658816.2012.698015.