Appendix S1: Statistical methods and additional figures and tables for

Integrating Multiple Data Types to Connect Ecological Theory and Data Among Levels

Jian D.L. Yen, Zeb Tonkin, Jarod Lyon, Wayne Koster, Adrian Kitchingman, Kasey Stamation, Peter A. Vesk

Corresponding author: Jian D.L. Yen, The University of Melbourne, Parkville, VIC 3010, Australia; P: +61 (3) 8344 4000; E: jian.yen@unimelb.edu.au

Statistical methods

Our aim was to estimate the parameters of a density-dependent Leslie matrix model:

$$\mathbf{n_{t,r}} = (1 \, / \, (1 + k_r \, N_{t-1,r})) \, \mathbf{A_r} \, \mathbf{n_{t-1,r}},$$

from data on size-abundance distributions, size-at-age, and binary recapture histories. Here, $\mathbf{n_{t,r}}$ is a vector containing the abundances of each age class at time t in river r and $\mathbf{A_r}$ is a matrix of vital rates (survival probabilities and fecundity estimates) in river r. $N_{t,r}$ is the total number of individuals in the population at time t in river r, and the parameter k_r governs the strength of density dependence in river r, with values close to zero indicating no density dependence and positive values indicating negative density dependence. We used five age classes and eight size classes based on the following bins: (0 g, 200 g], (200 g, 500 g], (500 g, 1000 g], (1000 g, 2000 g], (2000 g, 5000 g], (5000 g, 10000 g], (10000 g, 20000 g], and (20000 g, 60000 g]. These bins were chosen arbitrarily, with unequal bin widths to avoid the majority of individuals falling into one or a few size classes.

We used time series of size-abundance distributions to estimate vital rates, using size-at-age data to convert observed size-class abundances to ages, and using binary recapture histories to estimate the probability of detecting an individual in any given survey. We connected the population matrix model to the three data types with three component likelihoods:

Size-abundance distributions ($\mathcal{L}_{\text{abundance}}$): $\mathbf{y_{t,r}} \sim \text{Poisson}(p \, \mathbf{\Omega} \, \mathbf{n_{t,r}})$;

Capture histories ($\mathcal{L}_{\text{capture}}$): $\mathbf{z_i} \sim \text{CJS}(p, \mathbf{s})$; and

Size-at-age ($\mathcal{L}_{\text{size-age}}$): $\mathbf{u_i} \sim \text{Multinomial}(v_i, \mathbf{\omega_i})$.

The first component likelihood assumes that abundances in all size classes at time t in river r ($\mathbf{y_{t,r}}$) are independently Poisson-distributed, conditional on unobserved initial abundances ($\mathbf{n_{0,r}}$) and the matrix population model outlined above (i.e., age classes are connected through the Leslie matrix, $\mathbf{A_r}$). Observed abundances are reduced relative to true abundances due to

imperfect detection, with a size-independent detection probability p. Observed size-class abundances are converted to age-class abundances by the matrix $\mathbf{\Omega}$, which captures the probability that an individual in size class i belongs to age class j, for all i and j. The second component likelihood assumes that observed binary capture histories, $\mathbf{z_i}$, follow a Cormack-Jolly-Seber model with time-varying survival probabilities, $\mathbf{s}$, and constant detection probability, p (Lebreton et al. 1992). This use of capture histories allows estimation of the parameter p, used in the first component likelihood. The third component likelihood assumes that the distribution of individuals in size class i across all age classes follows a multinomial distribution with size v_i and probabilities $\mathbf{\omega_i}$. The size v_i is the number of individuals in size class i, and the data $\mathbf{u_i}$ is the observed distribution of individuals in size class i among all age classes. The probability vectors $\mathbf{\omega_i}$ formed the rows of the matrix $\mathbf{\Omega}$, which was used to convert modelled ages to size classes in the first component likelihood.

This model structure requires priors on the Leslie matrices ($\mathbf{A_r}$), the parameters p and $\mathbf{s}$ of the Cormack-Jolly-Seber model, the matrix $\mathbf{\Omega}$, age-class abundances at time 0 ($\mathbf{n_{0,r}}$), and the Beverton-Holt density-dependence parameter (k_r). We used a mixture of vague and vaguely informative priors, drawing on past empirical studies to inform estimates of survival and size-age associations. We did not assess sensitivity to choice of priors because our aim was to illustrate the implementation of a simple integrated model rather than present a rigorous analysis of our data.

Leslie matrices are sparse, with non-zero survival probabilities on the lower diagonal and on the diagonal in the final age class, which includes all individuals five years or older. We assigned inverse-logit transformed Gaussian priors to survival probabilities (hereafter, $\mathbf{p_{survival}}$), with priors in each river sharing a common mean and standard deviation. We assumed that mean survival was an increasing function of age:

$$\text{logit}(\mathbf{p_{survival,r}}) \sim \text{Normal}(\mathbf{\mu_{survival}}, \mathbf{\sigma_{survival}});$$

$$\boldsymbol{\mu_{survival}} = \alpha_{survival} + \beta_{survival} \times \textit{age}; \; \boldsymbol{\sigma_{survival}} \sim \text{HalfNormal}(0, 0.5);$$

$$\alpha_{survival} \sim \text{Normal}(0, 2); \; \beta_{survival} \sim \text{Normal}(0, 2).$$

Here, parameters in bold are vectors with five elements (one for each age class), and the subscript r denotes survival values in river r. The tilde is shorthand for "is distributed as", and all Normal distributions are parameterised with means and standard deviations. A HalfNormal distribution is a Normal distribution truncated to non-negative values.

We assumed that reproduction occurs from five years of age (Yen et al. 2013), so that the Leslie matrix included one non-zero element for reproduction (hereafter, $n_{offspring}$). We assigned HalfNormal priors to reproduction in each river, with a mean and standard deviation shared among all rivers:

$$n_{offspring,r} \sim \text{HalfNormal}(\mu_{offspring}, \sigma_{offspring});$$

$$\mu_{offspring} \sim \text{HalfNormal}(0, 3); \; \sigma_{offspring} \sim \text{HalfNormal}(0, 3).$$

All definitions follow those used for survival probabilities.

We used half-Gaussian priors to model the (unobserved) initial age-class abundances ($\mathbf{n_{0,r}}$) in river r, with a common mean and standard deviation among all rivers:

$$\mathbf{n_{0,r}} \sim \text{HalfNormal}(\boldsymbol{\mu_0}, \boldsymbol{\sigma_0});$$

$$\boldsymbol{\mu_0} \sim \text{HalfNormal}(0, 50); \; \boldsymbol{\sigma_0} \sim \text{HalfNormal}(0, 50).$$

Here, parameters in bold are vectors with one element for each age class. All other definitions follow those used for survival probabilities.

We used Dirichlet priors for the multinomial probabilities $\boldsymbol{\omega_i}$ in each size class i. These priors were designed to be informative given relatively few observations of fish with known age in larger size classes. Specifically, we set the Dirichlet concentration parameters for $\boldsymbol{\omega_i}$ to $80000 \times \exp(-\boldsymbol{\alpha_i}^2 / 2)$, where $\alpha_{i,j} = \{(10 / 7) \times (1 - i) + 2j\}$. This prior favours age-size associations that are relatively concentrated along the main diagonal of the age-size matrix $\boldsymbol{\Omega}$, so that larger fish are likely to be older than smaller fish.

Last, we assigned the survival probabilities (**s**) independent beta priors with both parameters equal to one and assigned the detection probability a beta prior with both parameters equal to 10. We set independent uniform priors on the density-dependence parameters in each river (k_r), with lower bounds of 10^{-5} and upper bounds of 0.2.

We assumed the three component likelihoods were independent, so that the composite likelihood was the product of all three component likelihoods:

$$\mathcal{L}_{composite} = \mathcal{L}_{growth} \times \mathcal{L}_{abundance} \times \mathcal{L}_{capture}.$$

Given this composite likelihood, we used to *greta* R package to generate fully Bayesian parameter estimates (Golding 2018). We based parameter estimates on 40000 random-walk Metropolis-Hastings Monte Carlo iterations (four chains of 100000 iterations, retaining every tenth sample), following a 1900000 iteration warm-up and burn-in period. This model took approximately 4 hours to run on a MacBook Pro with 2.5 GHz Intel Core i7 processor and 16 GB of RAM.

We assessed model convergence through visual inspection of chains and used Bayesian r^2 values to summarise model fit (Gabry & Goodrich 2018). Bayesian r^2 values extend the classical definition of r^2 (variance of predicted values divided by variance of the data) by replacing the denominator with the variance of predicted values plus the variance of the errors, which avoids situations in which the numerator exceeds the denominator (Gabry & Goodrich 2018). We did not use more-rigorous model validation (e.g., cross validation) because our goal was to illustrate the implementation of a simple integrated model rather than to present a full analysis of these data. Links to model code are in the *Data availability* section in the main text.

Table S1. Bayesian r^2 values between fitted and observed abundances. Bayesian r^2 values extend the classical definition of r^2 (variance of predicted values divided by variance of the data) by replacing the denominator with the variance of predicted values plus the variance of the errors, which avoids situations in which the numerator exceeds the denominator.

River system	Broken	Campaspe	Goulburn	Ovens	Murray	Loddon
Mean	0.48	0.46	0.38	0.38	0.02	0.42
Median	0.52	0.51	0.48	0.43	0.00	0.52
10% quantile	0.41	0.27	0.10	0.09	0.00	0.09
90% quantile	0.52	0.52	0.51	0.53	0.04	0.52

Table S2. Posterior mean, median, and 80 % credible intervals for the density dependence parameter k in each river system. Density dependence was modelled with a Beverton-Holt function with constant parameter k for all elements of the transition matrix. Values of k near zero indicate no apparent density dependence and values of k greater than zero indicate negative density dependence. The prior distribution of k did not allow negative values, so that positive density dependence was not included in fitted models. The effects of estimated k values on vital rates at different abundances are shown in Figure S2 (below).

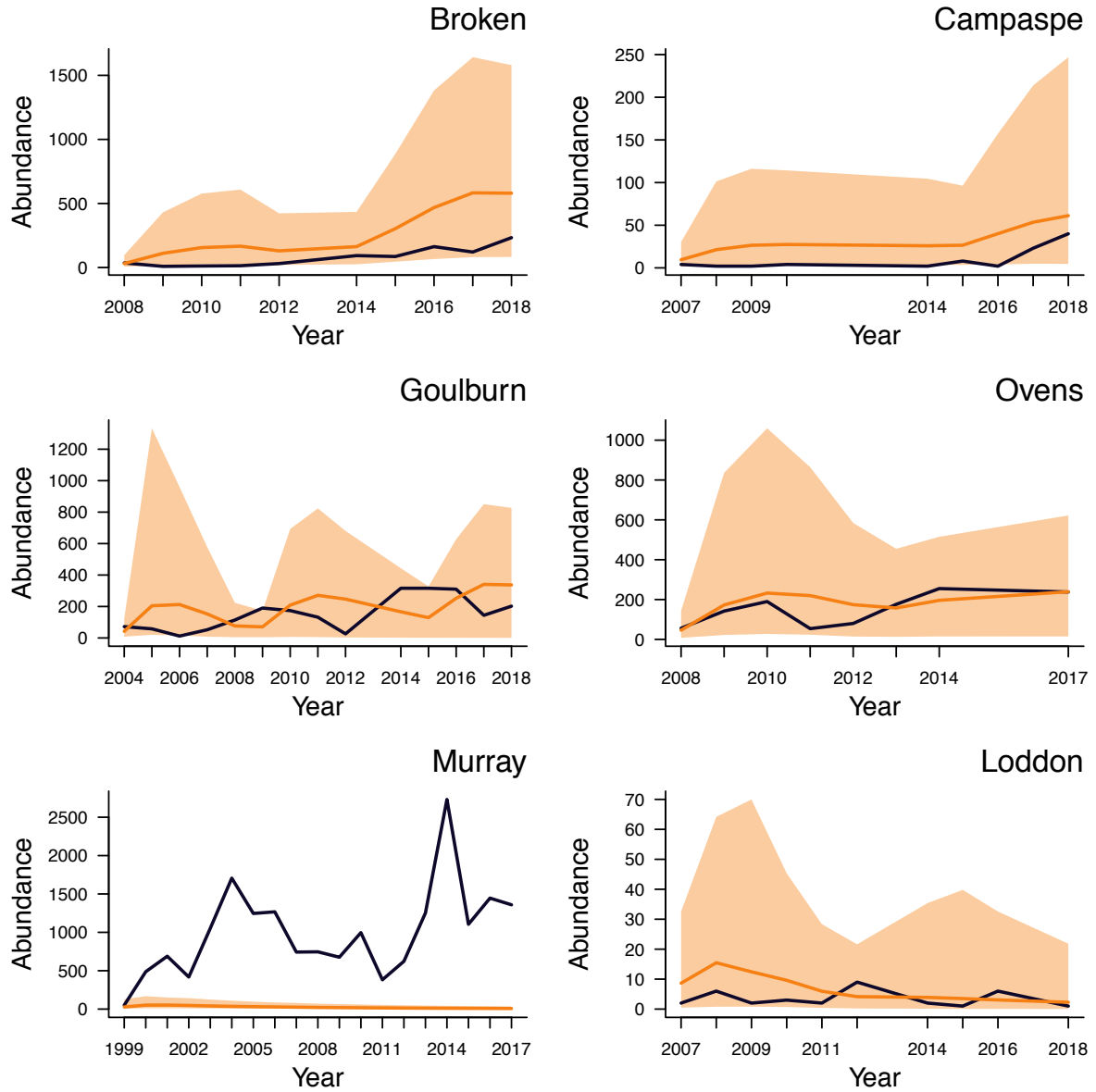| River system	Broken	Campaspe	Goulburn	Ovens	Murray	Loddon
Mean	0.00048	0.00187	0.00044	0.00032	0.00005	0.00862
Median	0.00005	0.00113	0.00001	0.00008	0.00002	0.00160
10% quantile	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
90% quantile	0.00029	0.00459	0.00118	0.00036	0.00009	0.02951

Figure S1. Observed abundances (grey) of Murray cod in each of six river systems plotted against estimated abundances from an integrated model (orange). Solid lines are means and shaded regions denote 2.5th and 97.5th quantiles.
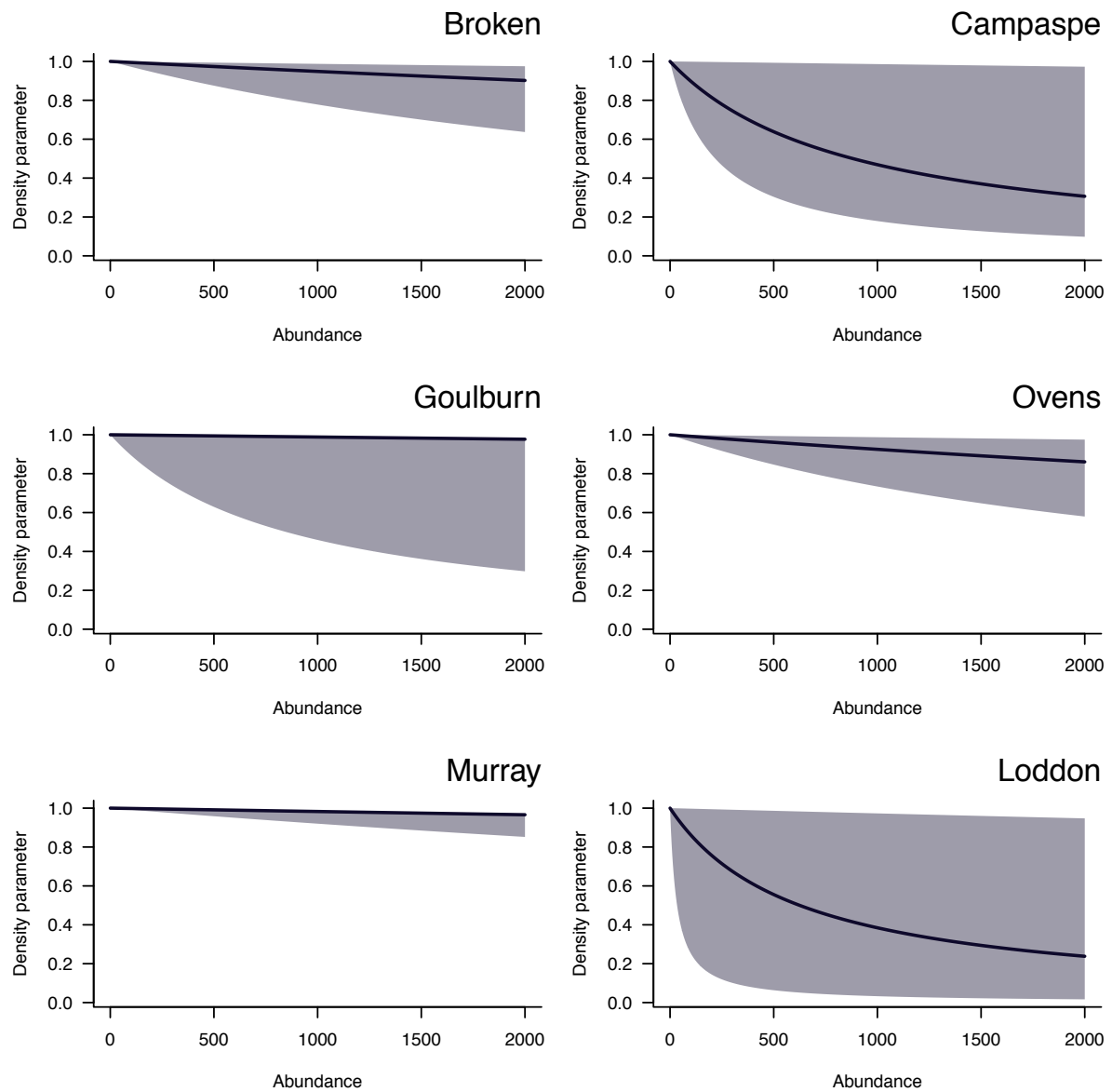
Figure S2. Fitted density dependence scaling factors for Murray cod in six river systems in south-eastern Australia. Scaling factors are the proportional reduction in all vital rates for a given abundance.