Is the integrated assessment the best place for modelling the growth of Macquarie Island toothfish?

R. M. Hillary, J. Day, and M. Haddon
CSIRO
Oceans & Atmosphere Flagship,
Castray Esplanade,
Hobart,
TAS 7000,
Australia.

Abstract

Over the last few years the case for estimating growth inside of integrated stock assessment models, using conditional age-at-length data, has been increasingly expounded. The first argument is that, quite rightly, this type of approach can solve the issues around bias in traditional length-at-age data that arise from length-specific selectivity. The second is that it can propagate the uncertainty in the growth parameters to the variables of interest in the assessment itself. Using extensive, sexually disaggregated multi-year ageing data for the Macquarie Island Patagonian toothfish (*Dissostichus eleginoides*) fishery we (i) compare (and select) traditional and hierarchical methods to modelling length-at-age outside the assessment; (ii) develop a novel method to analyse these data in the length-conditional framework, thereby by-passing the necessity of using the assessment; and (iii) compare the estimates to those obtained from the full assessment using all the available data. We obtained very consistent estimates for both the length-at-age and conditional age-at-length approaches, whereas the stock assessment estimates differed substantially for the males. The cause is the influence of non age-specific data in the assessment driving estimates that are at odds with the most informative growth data. This work highlights one of the main risks with modelling growth inside of the assessment: the analyst has no control over which data informs the parameter of interest, and strong potentially spurious trends in unexpected data sets may combine with structural issues in the model to dominate clearly informative data sets, resulting in anomalous parameter estimates. For the example in question we suggest solutions that balance the issue of biased estimates with the second - and still very important - issue of variance propagation to they key assessment variables.

Introduction

Growth is a central driver of the individual (and population-level) dynamics of fish (and the stock they belong to) as well as those of the fishery itself. As befits this centrality, modelling growth has been a mainstay of fisheries science and stock assessment from the beginning. Many sources of data hold information on growth.:

Direct length-at-age, where age is determined from a hard part such as an otolith or, for elasmobranchs, the vetertebrae. Mark-recapture data, where the growth increment between release and recapture is the information source, but there is no information on actual age. Length frequency data: in a closed population (specifically within one recruitment cycle) the progression of modes in length frequencies holds information on the growth rate of the population. As with mark-recapture data there is no direct information on age, save the number of age classes present if they are clear enough to be estimated well.

The focus of this paper is for cases where the primary growth data are length and age measurements and, more specifically, modelling these data both inside or outside of an integrated stock assessment (Methot and Wetzel, 2013). For Macquarie Island toothfish, there is an extensive amount of sexually disaggregated length-at-age measurements from otoliths so there is ample information available for exploring appropriate growth models. The stock assessment model itself is a modified version of the Stock Synthesis (Methot and Wetzel, 2013) modelling framework, and is sex-specific and spatially disaggregated, with fishery length-composition, conditional age-at-length, age-structured mark-recapture, and catch biomass data (Day *et al.*, 2015). The assessment is spatially structured, with two regions where age-independent movement is estimated within the assessment and the iterative reweighting approach outlined in (Francis, 2011) is used to weight the various data sets. The structure of the assessment has morphed over time as more and more data have become available, and early issues around the estimation of all the parameters of the female growth curve (Fay , 2010) led to asymptotic length becoming a fixed parameter in the assessment henceforth. The first indication of a potential problem appearing with the males was observed in 2014 (Day *et al.*, 2014), where the growth rate and maximum length parameters were suddenly estimated to be much smaller and larger, respectively, than both previous estimates and their female counterparts, at apparent odds with the age data that clearly showed males being consistently shorter than females (Day *et al.*, 2014). This problem worsened in 2015 with nonsensical estimates of both growth rate and asymptotic length and a notable deterioration in the actual fitting to the age data over the sample range (Day *et al.*, 2015). The main question was this: are the differences in the estimates both inside and outside the assessment driven by different assumptions and probability model structures, or something being driven by the assessment itself and the other data therein?

Estimating growth via paired length-at-age measurements has been the most common approach, historically speaking, presumably because of its relative ease of implementation. It assumes that length is random for a given age; age is known without error; and each length observation is conditionally independent of the others, given the model parameters. In almost all cases, and certainly for the specific example of Macquarie Island toothfish, we are fairly certain the selectivity is driven by length (driven by depth-specific ontogenic shifts as animals grow), we know that the age estimates are subject to error (Candy *et al.*, 2012), and the length measurements are almost certainly not independent.

More recently, alternate approaches have explored the age-at-length framework (Morton & Bravington, 2008; Piner *et al.*, 2015). In this approach the primary data source is the relative frequency of age *conditional* on length, not length-at-age. This assumes that age is sampled randomly within the specific length bin of interest. In the integrated stock assessment framework (Methot and Wetzel, 2013) the sample bias, attributed to length-specific selectivity, is dealt with directly as the selectivity is estimated along with the growth parameters via the inclusion of the length frequency data. Additionally, the ageing error can also be dealt with efficiently in this framework. The disadvantages of this approach are that it can be harder to visualise the data, relative to the length-at-age framework, and it is computationally more demanding. An additional disadvantage to implementing this approach in the integrated assessment framework is parameter aliasing and correlation with other (mis-specified) life-history parameters, as well as potentially spurious information from other data sources like abundance indices.

In this paper we explore both approaches to modelling growth, compare and contrast the estimates using previous growth modelling work for this stock and a novel age-at-length approach that can be undertaken outside of the assessment framework. These external estimates are then compared to those from the stock assessment. The point of the external age-at-length model is to serve as a bridge to the full assessment estimates. Length-at-age estimates have often differed from the age-at-length estimates in the assessment model - particularly for the males (Day *et al.*, 2014, 2015). By having the external age-at-length model we can determine the cause of those differences: is it the alternative modelling approach, or the interaction with other data and life-history parameters in the assessment? We also suggest an approach to balance unbiased parameter estimation with variance propagation in the stock assessment.

Materials and methods

Data

The length-at-age data, for which we have sex identification, are quite extensive: 1,921 males and 2,756 females, caught between 1996 and 2015. For some fish there are repeated measurements, and in such cases the age used is the mean age of the repeated readings. Ages taken from recaptured tagged fish are also excluded in certain runs given potential concerns around post-tag growth retardation leading to biased lengths for those fish. Figure 1 summarises the length-at-age data used in these analyses.

Exactly the same underlying data are used in the age-at-length analyses, but they require a reasonable amount of transformation first. We need to define a length partition so as to be able to bin the data and we chose the following: 0.2m defines the lower bound of the partition (for both sexes) and the partition is split into equal sized bins of width 0.1m up to 1.5m and the final length bin stretches from 1.5m to the maximum observed length of that particular sex. The two key data sets are:

1. The length frequency distribution, $p_{y,l}$, of the aged animals in year y
2. The age frequency distribution, $p_{y,a,l}$, *conditional* on the given length partition element l, in year y

The summation conventions with the two frequency distribution data sets are as follows: $\sum_l p_{y,l} = 1$ and $\sum_a p_{y,a,l} = 1$. Figure 2 shows the length distribution by both year and sex.

Length-at-age

As mentioned, the base model is von Bertalanffy with a normal-log likelihood function. This is very similar to a normal likelihood with errors proportional to length which, at least for the process error term, is one of the ways to model individual variation in L_∞ from a non-hierarchical perspective. So $\widehat{l} = L_\infty \left(1 - e^{-k(a-t_0)}\right)$ and $\log l \sim N(\log \widehat{l}, \sigma_o^2 + \sigma_p^2)$, $\sigma_o^2 = \log(1 + cv_o^2)$ is the observation error (for the given CV of 0.05 for this case), and σ_p^2 is an estimated process error term. The estimated parameters are $\ell_\infty = \log L_\infty$, $\kappa = \log k$, t_0 and σ_p^2 with the following priors:

$$
\begin{aligned}
\ell_\infty &\sim N(\mu_{\ell_\infty}, \sigma^2_{\ell_\infty}), \\
\kappa &\sim N(\mu_\kappa, \sigma^2_\kappa), \\
t_0 &\sim N(\mu_{t_0}, \sigma^2_{t_0}), \\
\sigma_p &\sim \sigma_p^{-1},
\end{aligned}
$$

with $\mu_\bullet = 0$ and $\sigma^2_\bullet = 100$ in all cases. With the assumption of a normal likelihood on the log-scale length data the conditional posterior for ℓ_∞ is known:

$$\pi(\ell_\infty \mid \cdots) = N\left(\tilde{\mu}, \tilde{\sigma}^2\right) \tag{1}$$

where

$$\tilde{\mu} = \left(\frac{\mu_{\ell_\infty}}{\sigma^2_{\ell_\infty}} + \frac{\sum_{i=1}^{\aleph} \varepsilon_i}{\sigma^2_o + \sigma^2_p}\right)\left(\frac{1}{\sigma^2_{\ell_\infty}} + \frac{\aleph}{\sigma^2_o + \sigma^2_p}\right)^{-1},$$

$$\tilde{\sigma}^2 = \left(\frac{1}{\sigma^2_{\ell_\infty}} + \frac{\aleph}{\sigma^2_o + \sigma^2_p}\right)^{-1},$$

and $\aleph$ is the number of data points and $\varepsilon_i = \log l_i - \log\left(1 - \exp(-k(a_i - t_0))\right)$. This makes it a bit more efficient in terms of MCMC sampler performance, as ℓ_∞ is sampled directly from the conditional posterior in (1) and κ, t_0 and σ_p are updated using a Metropolis-within-Gibbs routine written up in `C++`. A 1,000 iteration burn-in is used, with a further 1,000,000 samples drawn and every 1000^{th} retained (thinning factor of 1000) to yield 1,000 samples from the posterior (non-convergence checked using standard methods (Brooks and Roberts, 1998)). The reason for the high level of thinning is the strong levels of Markov chain autocorrelation you get with such large and informative data sets, and very correlated parameters like k and L_∞.

Individual asymptotic length

The hierarchical model is a little different in form than the previous example. The first difference is that we remove the process error form from the likelihood, so $\widehat{l}_i = L_{\infty,i}\left(1 - e^{-k(a_i - t_0)}\right)$ and $\log l_i \sim N(\log \widehat{l}_i, \sigma^2_o)$ and the index i relates to each animal that is aged and measured: each animal has its own L_∞. In this formulation, we assume that $\mu_{\ell_\infty} \sim N(\mu_\xi, \sigma^2_\xi)$ and $\sigma^2_{\ell_\infty} \sim IG(\gamma_\xi, \psi_\xi)$ (inverse gamma distribution) and these are our hyper-parameters with the following hyper-priors: $\mu_\xi = 0$, $\sigma^2_\xi = 100$, $\gamma_\xi = \psi_\xi = 0.001$. This model has *considerably* more parameters than the previous one (for the females/males 1,926/2,761 versus 4). Fortunately, the likelihood and prior structure means that the conditional posteriors for $l_{\infty,i}$, μ_{ℓ_∞} and $\sigma^2_{\ell_\infty}$ are all of a known form:

For $l_{\infty,i}$, we have the following:

$$\pi(l_{\infty,i} \mid \cdots) = N\left(\left(\frac{\mu_{\ell_\infty}}{\sigma^2_{\ell_\infty}} + \frac{\varepsilon_i}{\sigma^2_o}\right)\left(\frac{1}{\sigma^2_{\ell_\infty}} + \frac{1}{\sigma^2_o}\right)^{-1}, \left(\frac{1}{\sigma^2_{\ell_\infty}} + \frac{1}{\sigma^2_o}\right)^{-1}\right), \tag{2}$$

and for μ_{ℓ_∞}:

$$\pi(\mu_{\ell_\infty} \mid \cdots) = N\left(\left(\frac{\mu_\xi}{\sigma_\xi^2} + \frac{\sum_{i=1}^{\aleph} l_{\infty,i}}{\sigma_{\ell_\infty}^2}\right)\left(\frac{1}{\sigma_\xi^2} + \frac{\aleph}{\sigma_{\ell_\infty}^2}\right)^{-1}, \left(\frac{1}{\sigma_\xi^2} + \frac{\aleph}{\sigma_{\ell_\infty}^2}\right)^{-1}\right), \tag{3}$$

and finally for $\sigma_{\ell_\infty}^2$:

$$\pi(\sigma_{\ell_\infty}^2 \mid \cdots) = IG\left(\gamma_\xi + \frac{\aleph}{2}, \left(1 + \frac{\psi_\xi \sum_{i=1}^{\aleph} (l_{\infty_i} - \mu_{\ell_\infty})^2}{2}\right)\psi_\xi^{-1}\right) \tag{4}$$

Age-at-length

This approach makes fundamental use of the same assumed distribution for length-at-age as in the more traditional growth framework. For a given age, a, there is an associated probability of each length bin in the partition, l:

$$\pi_{l|a} = CDF\left(\mu_{\lceil l \rceil,a}, \sigma_{\lceil l \rceil,a}\right) - CDF\left(\mu_{\lfloor l \rfloor,a}, \sigma_{\lfloor l \rfloor,a}\right), \tag{5}$$

where $CDF()$ is the Gaussian cumulative distribution function; $\mu_\bullet$ and $\sigma_\bullet$ are the log-scale mean length and associated SD, respectively; and $\lfloor l \rfloor$ and $\lceil l \rceil$ are the infimum and extremum of the partition element l, respectively.

To get the (annual) probability distribution of age-at-length, $\pi_{y,a|l}$, we must apply Bayes' rule:

$$\pi_{y,a|l} = \frac{\pi_{l|a}\pi_{y,a}}{\pi_{y,l}}, \tag{6}$$

where

$$\pi_{y,l} = \sum_a \pi_{l|a}\pi_{y,a}. \tag{7}$$

It is at this stage that we need to define additional parameters - specifically the prior age-distribution, $\pi_{y,a}$. In a simple equilibrium population with constant, age-independent natural (M) and fishing (F) mortality and no recruitment variation the age distribution in the population is proportional to $\exp(-Za)$, where $Z = F + M$. Indeed, this forms the basis of the approximation used in Piner *et al.* (2015) to avoid having to explicitly model the prior age distribution. In the stock assessment context, with a simple time-invariant selectivity ogive, s_a, then this prior is proportional to $s_a N_{y,a}$, where $N_{y,a}$ is the numbers-at-age matrix.

We, however, take a more direct parametric route and model each year-specific prior age distribution as a log-normal distribution, with estimated mean μ_y and SD σ_y.

The conditional age-at-length data alone are insufficient to estimate both the growth and the prior age distribution parameters, but the inclusion of the length frequency data *can* make joint estimation feasible. We say can make it feasible because that depends on the amount of information on the likely age distribution in the sample contained within the length data. If all the length data are too close to where growth begins to asymptote the amount of information on age dramatically reduces. Additionally, there is often an increase in the precision of the growth estimates that accompanies the inclusion of the length frequency data (Morton & Bravington, 2008). As we see in Figure 2 the length frequencies of the aged samples are (mostly) unimodal and slightly right-skewed (hence, the choice of a log-normal age prior) and not clustered at the highest observed lengths and so are likely to be informative with respect to the prior age distribution.

The basic (multinomial) log-likelihood for the conditional age-at-length data is as follows:

$$\Lambda_{a|l} \propto \sum_y \sum_l \sum_a n_{y,l} p_{y,a,l} \log\left(\pi_{y,a|l}\right), \tag{8}$$

and for the length data (also multinomial)

$$\Lambda_l \propto \sum_y \sum_l n_y p_{y,l} \log\left(\pi_{y,l}\right), \tag{9}$$

and n_y is the number of samples of lengthed and aged fish taken in that year, and $n_{y,l} = n_y p_{y,l}$. The total log-likelihood, Λ, is just the sum of the two terms in Eqns. (8) & (9).

Ageing error

Ageing error is a major factor in general for estimating age from hard parts. Repeat readings of otoliths across different readers enables the construction of an ageing error matrix, $A_{a,a'}$, the $\{a, a'\}^{\text{th}}$ element of which is the probability that the observed age a is actually the "true" age a', and $\sum_a A_{a,a'} = 1$. Essentially, each of column of the matrix defines $\Pr(a \,|\, a')$, and the model-prediction for the observed proportion of age a in the sample, is given by

$$\pi_{y,a|l} = \sum_{a'} \Pr(a \,|\, a') \pi_{y,a'|l}, \tag{10}$$

where $\pi_{y,a'|l}$ is the model-predicted proportion of actual age a' fish in the sample.

The current ageing error matrix is based on the Heard and McDonald Island toothfish fishery's ageing data (Candy *et al.*, 2012) and can be visualised in Figure 3. Ideally an

ageing error matrix would be calculated for the Macquarie Island data specifically, but given the same readers process the Macquarie Island data as read the data used in (Candy *et al.*, 2012), and they are the same species with very similar apparent life histories, we feel this assumption is valid. Accommodating ageing error in the length-at-age paradigm is technically quite difficult, requiring a very high dimensional state-space augmentation as it becomes an errors-in-variables problem. In the age-at-length framework, it can be fully accounted for in the likelihood as defined in Eq. (8).

Process error

One key assumption for both the age-at-length and the length data is that the number of (paired) length and aged samples, n_y, are the true number of independent samples - the *effective* sample size. Another is that the true level of variability in the length-at-age relationship and prior age distribution is captured in the probability model defined. The first assumption is very often invalid as a lot of sampling is - to some degree - not randomly done; the second assumption can be broken by a mis-specified model in the variation in length-at-age (e.g. "fatter" tails than the Gaussian model assumed). Given the underlying multinomial likelihood the natural way to do this is via the compound Dirichlet-multinomial distribution. This model assumes a Dirichlet distribution for underlying model-predicted probability $\pi_\bullet$ (be it age-at-length or just length). The key parameter that controls this distribution is ω.

In a very general setting with data Y, model-predicted sampling probability, π, and underlying sampling probability ξ - where $\xi \sim p(\xi \,|\, \pi, \omega)$ - the joint likelihood is defined as:

$$\ell(Y \,|\, \xi, \pi, \omega) \propto \ell(Y \,|\, \xi) p(\xi \,|\, \pi, \omega). \tag{11}$$

What we need to do is integrate over ξ to leave the marginal likelihood of the data given the model predicted sampling probability:

$$\ell(Y \,|\, \pi, \omega) \propto \int \ell(Y \,|\, \xi) p(\xi \,|\, \pi, \omega) \mathrm{d}\xi. \tag{12}$$

This integral is of a known form, and the marginal log-likelihood for the length data can be expressed interms of the log-scale gamma function $\gamma = \ln \Gamma()$:

$$\Lambda_l(n_y, p_{y,l} \,|\, \pi_\bullet, \omega_\bullet) \propto \sum_y \left(\gamma(n_y + \omega_y) - \gamma(\omega_y) + \sum_l \gamma\left(\omega_y \pi_{y,l}\right) - \gamma\left(n_y p_{y,l} + \omega_y \pi_{y,l}\right) \right). \tag{13}$$

For the conditional age-at-length data it is a bit more involved, given the more complex nature of the data:

$$\Lambda_{a|l}(n_y, p_{y,l,a} \,|\, \pi_\bullet, \omega_\bullet) \propto \sum_y \sum_l \left(\gamma(n_y p_{y,l} + \omega_{y,l}) - \gamma(\omega_{y,l}) + \sum_a \left[\gamma\left(\omega_{y,l}\pi_{y,a|l}\right) - \gamma\left(n_{y,l}p_{y,l,a} + \omega_{y,l}\pi_{y,a|l}\right) \right] \right). \tag{14}$$

In the multinomial-Dirichlet formulation, the key additional variance parameter, ω, is better understood (and defined) via the concept of an over-dispersion coefficient, φ. With sampling probability π and sample size n the mean and variance of the multinomial are $n\pi$ and $n\pi(1-\pi)$, respectively. When assuming the compound multinomial-Dirichlet model then mean is unchanged, but the variance is now $\varphi n\pi(1-\pi)$ where

$$\varphi = \frac{n+\omega}{\omega+1}, \tag{15}$$

which, if one is controlling the additional variance via φ or are estimating this parameter directly, leads to the following simple equation for ω:

$$\omega = \frac{n-\varphi}{\varphi-1}. \tag{16}$$

A note of caution when using this formulation is $\varphi \to 1$. In the limiting case, the true sampling distribution ξ approaches a point distribution (i.e. zero variance). While theoretically valid, numerically it can be very problematic, so if one is estimating φ, make sure it is penalised to stay away from one.

In this paper, given the fairly large number of over-dispersion parameters contained in both the likelihoods, an empirical Bayes approach (Casella , 1985) is used to estimate them. In this approach one estimates the most likely values of the hyperparameters (the $\varphi_\bullet$ in particular) which maximise the marginal likelihood of the data. For the MCMC analyses they are kept fixed at their respective MMLEs, which ignores the uncertainty in the estimates themselves. Given we use a predictive approach to model performance, any issues arising from this choice should become apparent. An advantage of this formulation is that one never *up-weights* the data, which can happen in assessment contexts when the underlying initial sample size, n, is either unknown or not really believed.

Additional priors and MCMC algorithm

The conditional age-at-length model requires the estimation of the four key growth parameters (L_∞, k, t_0, and σ_p) but also the mean (μ_y) and standard deviation (σ_y) of the prior age distribution $\pi_{y,a}$ for each year for which we have length and age data. The prior mean across all years was just the average age across the observed range. The prior SD was chosen so that the observed age range was contained within the prior 95% probability interval.

Given there are a moderately high number of estimated parameters now (68) and no obvious conjugacies between the likelihood and prior distributions, which made the hierarchical length-at-age models tractable despite their massive number of parameters, a different MCMC approach is used. We implemented what is called Hamiltonian MCMC (Neal, 2011) to sample from the full posterior distribution. This is a very powerful more recent MCMC suite of algorithms that is particularly useful when sampling from large dimensional distributions with highly correlated parameters, just as we must. The specifics of the scheme we implemented can be found in the Appendix, but using this approach we were able to obtain very high acceptance rates for new sample proposals while maintaining very good mixing characteristics for the Markov chains. The end result is that we required far fewer overall iterations of the scheme - 10,000 to achieve a convergent sample of 1,000 posterior draws versus 1,000,000 before - and commensurately lower run times.

Model performance criteria

In the length-at-age framework, given the simpler nature of the model there are a number of powerful MCMC tools we can use to analyse how well the model is performing in terms of predicting the observed data. When comparing the two candidate length-at-age models - simple vs. hierarchical - posterior predictive analysis (Gelman *et al.*, 2006; Meng, 1994) is both relatively simple to implement and very informative. This is done as follows:

1. For a given posterior sample, length-at-age $\tilde{l}$ data are simulated from the likelihood
2. The simulated residual $\tilde{l} - \widehat{l}$ is calculated
3. The observed residual $l - \widehat{l}$ is also calculated
4. The absolute median deviation in each of these 1,000 residuals, $\tilde{\Delta}$ and Δ, is calculated
5. The statistic $p(\tilde{\Delta} > \Delta)$, known as a Bayesian p-value, is calculated and the plot of the predicted versus observed discrepancy statistics is also useful

The main idea is that if $p(\tilde{\Delta} > \Delta) > 0.5$ the predictions are generally more variable than the observations; *vice versa* for $p(\tilde{\Delta} > \Delta) < 0.5$. Finally, p-values outside the range of 0.05-0.95 are generally indicative of *some* kind of issue with the model and/or likelihood (Gelman *et al.*, 2006).

For the age-at-length data, the whole concept of model performance and predictive interpretation is more complicated than the length-at-age data. For a start, we have two predicted quantities not one: length distribution of the aged samples *and* their associated conditional

age-at-length distributions. In the case of the predicted length frequency this is, in effect, a matrix by both year and length partition element; for the age-at-length data it is a three-dimensional array of year/length/age. Unlike the underlying residual between predicted and observed length-at-age, there is no immediately obvious discrepancy statistic with which to perform a posterior predictive analysis.

For the length data, we have the expected proportion of each length-class by year, $\pi_{y,l}$. For each MCMC iteration and we can simulate a prediction, from the compound multinomial-Dirichlet likelihood, of the observed length composition $\tilde{p}_{y,l}$. We also obviously have the actual observed length composition, $p_{y,l}$. To construct a suitable - and univariate - discrepancy statistic we calculate the following two (residual) matrices:

$$\tilde{X} = \tilde{p}_{y,l} - \pi_{y,l}, \tag{17}$$

$$X = p_{y,l} - \pi_{y,l}. \tag{18}$$

Clearly, we need to some reduce the dimensionality of the matrices $\tilde{X}$ and X and we do this using the Frobenius matrix norm, $\|\|_F$:

$$\| M \|_F = \sqrt{\sum_i \sum_j m_{ij}^2} = \sqrt{\text{trace}\,(M^\dagger M)}, \tag{19}$$

where $\dagger$ denotes the matrix transpose. We now define the discrepancy statistics as $\tilde{\Delta} = \| \tilde{X} \|_F$ and $\Delta = \| X \|_F$, and our Bayesian p-value is the same as for the length-at-age data: $p(\tilde{\Delta} > \Delta)$.

For the conditional age-at-length data, we decided to focus on the age-at-length matrix in each year when constructing the discrepancy matrix:

$$\tilde{X}_y = \tilde{p}_{y,l,a} - \pi_{y,a|l}, \tag{20}$$

$$X_y = p_{y,l,a} - \pi_{y,a|l}. \tag{21}$$

The calculation is essentially the same as for the length data discrepancy, but now we have one for each year. The philosophical reasoning behind this is as follows:

For the length data, we are interested in capturing the general form of the distribution over the years, so eschew a year-specific discrepancy statistic. For the age-at-length data there are more processes at work that could lead to both year and length-specific process errors (hence the more detailed process error model for these data). It is for this reason that - for each year - we focus on the predictive performance of the probability model with respect to the observed age-at-length distribution.

Results

Initially, we address the results of the two different approaches separately, then try to compare and contrast them later on.

Length-at-age data

In terms of the visual fit to the data, Figure 4 summarises the posterior median and 95% predictive credible interval for the length-at-age of both sexes. Overall, there is no substantial difference between the basic and hierarchical visual data fits. For the females the hierarchical model predicts a very slightly wider predictive interval with very similar medians; for the males the hierarchical model predicts slightly lower length-at-age in general. Both seem to cover the spread of data reasonably well.

Figure 5 summarises the predictive performance of both the basic and hierarchical models for the female and male length-at-age data. In terms of Bayesian p-values, for the basic model the values were 0.69 and 0.97 for the female and male data, respectively; for the hierarchical model, they were 0.54 and 0.55. For both sexes, the basic model predicts higher variability in the discrepancy statistic than in the actual observations - particularly for the male data. Additionally the predicted discrepancy statistics show a much broader range than for the observed data. With the hierarchical model the p-values are very close to 0.5 (the "ideal" level) with a very symmetric spread so very consistent with the observations, and with actual values almost three times smaller than the basic model. From the predictive perspective, the hierarchical model outperforms the basic one on every level: p-value, symmetry in the discrepancy statistic and a universally better fit to the data.

In one sense, it should be better given it has vastly more parameters. Exactly how many free parameters a hierarchical model has is a complicated concept, given the estimated priors for L_∞ constrain the individual values. This, along with other factors, makes information criterion-based model selection approaches very complicated in the Bayesian hierarchical framework (Celeux *et al.*, 2006). One interpretation of the Deviance Information Criterion (DIC) (Gelman *et al.*, 2006) we calculated clearly favours the hierarchical model, and this statistic accounts for the additional parameters. However, we do not go further into this analysis as there are other interpretations (Celeux *et al.*, 2006) and it does not contradict the posterior predictive findings: the hierarchical model is the better choice for the length-at-age data for both sexes.

Age-at-length data

Figures 6 and 7 visually summarise the fits to the female and male length and age-at-length data, respectively, at the posterior median. The length data are, in general, reasonably well fitted to. The only notable issue is that model fails to capture the apparent bimodality in the female length frequencies in 2014 and 2015. This is not surprising, given we assume a unimodal lognormal distribution for the year-specific prior age distribution in the catch. The underlying (expected) age-at-length relationship is monotonically increasing, so no aspect of the model could capture the apparent bimodality.

For the fits to the conditional age-at-length data, we use the posterior median and 95% credible interval to predict the mean age in each length bin. For the females, the distribution of mean age-at-length appears well captured, with none of the data appearing outside the predicted credible intervals. For the males, this also seems to be the case apart from 2013, where the mean age at the lowest length bin is significantly higher than model would predict, and much lower for the 1.5–1.6m length bin. In terms of potential year-to-year variation in growth the years 1998 and 2004 seem to exhibit small but systematic differences in mean age-at-length for both sexes, with 2009 being a candidate for the males but not the females.

For these data we undertook predictive analses for the length and age-at-length data - for the former there was one p-value, and for the latter there was one for each year. For the length data, the Bayesian p-values for the females and males were 0.38 and 0.44, respectively. Both suggest that we are slightly under-estimating the true variability in the data, and more so for the females than the males (driven largely by the 2013 and 2014 discrepancies). For the conditional age-at-length data the year-specific values range from 0.33–0.58 and a median of 0.47 for the females, and from 0.37–0.56 with a median of 0.45 for the males. So, in general, we are *slightly* under-estimating the variability in the age-at-length data, but otherwise predicting these data well.

To emphasise the need for the detailed process error model we employed, when performing these analyses without any process-error reweighting (i.e. assuming the straight multinomial is correct) the predictive analyses all show often serious under-estimation of variability in both data sets. We estimated p-values to be mostly very low (less than 0.1) and often well below 0.05 - the point at which the general advice is something in your model is wrong. To sufficiently explain the data and, perhaps more importantly, to increase the likelihood of getting unbiased parameter estimates and not under-estimating parameter variances it may often require more complex probability models than are currently available in major stock assessment packages like Stock Synthesis (Methot and Wetzel, 2013).

Across all models

Table 1 summarises the key growth parameter estimates across all three model structures and for both the sexes. When comparing the two length-at-age approaches (basic and hierarchical) there is little consistent difference between the parameter estimates - in terms of both median and credible intervals. The only difference is for the estimates of female L_∞: for the hierarchical model it is 7% higher than for the basic model. All the other estimates are very close and all sit comfortably within their counterparts' 95% credible interval. One cannot directly compare the variability in L_∞ for the basic and hierarchical models: the basic model estimates one value assuming them to all share this parameter; the hierarchical estimates one for each animal *and* the population mean and variance (summarised in the table via the posterior predictive distribution for L_∞). A rough rule-of-thumb is to compare the CV for the hierarchical model with the square root of the sum of the squares of the CV for L_∞ and $CV_p = \log(1 + \sigma_p^2)$, as this parameter is trying to capture the underlying variation in individual L_∞. For both males and females this "effective" CV in population-level L_∞ inferred from the basic model is around 0.15, so very similar to that estimated more formally in the hierarchical model.

For simplicity, and given the similarity in the length-at-age estimates, we directly compare only the hierarchical length-at-age estimates with the conditional age-at-length estimates. In terms of posterior median estimates: L_∞ was estimated to be lower for the age-at-length approach for both males and females, though not significantly so; estimates of k were very similar for both sexes; t_0 was estimated to be closer to zero for the age-at-length approach, though not significantly so; and the estimates of σ_p were practically identical. In terms of variance the 95% credible intervals were larger when using the age-at-length approach - most notably for L_∞ and t_0.

There are no obvious model selection tests we could use to decide between the hierarchical length-at-age and conditional age-at-length approaches. For one thing, they are fundamentally different data sets, even though one is effectively derived from the other, so no information theoretic tests could be applied. From the posterior predictive analyses, the hierarchical length-at-age model performs very well; the conditional age-at-length model performs well, with a suggestion of under-estimation of variability in the data, though nothing close to being problematic.

Comparison with the full stock assessment growth estimates

In the current integrated assessment (Day *et al.*, 2015) the only fixed growth parameter is L_∞ for the females (at 1.65m) - all other growth parameters are estimated. In the 2014

assessment the estimate of L_∞ for the males was noticeably higher for the males (*ca.* 2m and above) than for any of the external estimates of growth (see Table 1), and given the negative correlation the estimates of k lower. In 2015 we saw the same effect though much increased: estimates of L_∞ and k were 18.5m and 0.003, respectively. In the stock assessment model the actual estimated growth parameters are k and the length-at-age for pre-specified ages a_1 and a_2 - L_∞ and t_0 are easily derived from these three parameters but not directly estimated. Figure 8 shows the predictive intervals for the paired length-age data for the length-at-age, conditional age-at-length, and stock assessment growth curves. For the males, one can clearly see that this is not just an issue of parameterisation (where L_∞ and k are "odd" but still fit the data in the sampled range). The assessment curve fundamentally fails to explain the length-at-age data even in the most densely sampled length intervals and ages. And we know this is not a selectivity-driven bias issue in the data themselves, as we get very consistent estimates even when accounting for this in the external length-conditional approach. The key question is: what data are driving the growth estimates within the stock assessment?

Table 2 details the most recent (Day *et al.*, 2015) stock assessment objective function breakdown for each data source and penalty term for four male L_∞ scenarios: 1.3m, 1.65m, 2m and the estimated value of 18.51m. For each element of the objective function the minimum value is subtracted, so a value of zero indicates the most preferred value, with increasing positive values indicative of an increasing lack of preference by that data set/penalty. In total, obviously the preference is for the estimated value, with the least preference for the externally estimated value of around 1.3m.

Clearly from Table 2 it is the tag data, both in terms of total recaptures and their spatial composition, that is causing the mismatch. The length and age-at-length data both support values less than 2m, whereas both the tagging data sets (total recaptures and spatial distribution thereof) strongly push for values well above this. The most likely candidate driver of this tag data driven push for higher values is age-specific movement. The current model assumes age-independent movement and the tag releases are focussed more on the smaller animals caught in the fishery, mostly to maintain very high survivability post-release. These animals are located in the length ranges where selectivity is estimated to be still increasing with length, not asymptoting or decreasing. If there is a tendency for movement to increase with length/age in this range of tagged animals then a model which assumes age independent movement is likely to result in biased predicted recapture data. Specifically in this case, it is quite likely to over-estimate the number of both total recaptures and their spatial distribution, as the model is moving too many fish around relative to what the actual data are saying. Given selectivity is clearly increasing in this tagged length range (and for a number of years post-release) one way the model can modify both overall tag return numbers and their spatial dispersal is to slow down growth. By having slower growth, the tagged fish

will ascend the selectivity curve slower and be less available to be caught by the fishery and result in fewer predicted recaptures. Looking at Figure 8 that is precisely what seems to be happening for both the female and the male growth curves - more so for the males given the asymptotic length is not fixed, but clearly mean length-at-age and growth rate are both lower than estimated from the data used outside of the assessment.

Importance of variance propagation in the assessment

One obvious answer to the recent divergence of the internally estimated growth parameters in the assessment is to simply fix the parameters at their externally estimated values, given their clear consistency across methods. This does, however, then raise a problem in the assessment and in particular management advice context. The variability in the growth relationship will be removed from the assessment results, and the setting of the total allowable catch (TAC) for this fishery explicitly takes into account the uncertainty in the spawning stock biomass projections (Day *et al.*, 2015). Given the specifics of the harvest control rule, the more/less uncertain the future SSB relative to the unfished state, the lower/higher the resultant TAC. Knowingly reducing the variability in the key management variable will result in (relatively) higher TACs, assuming relatively unbiased estimates in the first place.

So fixing all the growth parameters would, on face value, probably be upwardly biasing the TAC. The assessment is not currently run with all the growth parameters fixed so we cannot assess the degree to which growth variability contributes to overall variability in SSB depletion. We can, however, explore a more tractable problem and assess the relative influence of fixing only *some* of the growth parameters, with respect to overall variance in the management variables.

To explore how variability in growth propagates into the variables we are interested in for assessment purposes, let us consider the variation in SSB-per-recruit in the unfished state - this is a major contributor to the variation in B_0 (the unfished SSB), when R_0 (the unfished recruitment level) is the estimated parameter as in Stock Synthesis. For the unfished SSB-per-recruit, $SPR_{F=0}$, we need to define the equilibrium age structure: $\tilde{n}_a$. For $a = 1$, $\tilde{n}_a = 1$, and for $a = 2, ..., A-1$:

$$\tilde{n}_a = \tilde{n}_{a-1} \exp(-M), \tag{22}$$

and assuming a plus group at the maximum age A, we have that

$$\tilde{n}_A = \tilde{n}_{A-1} \frac{\exp(-M)}{1 - \exp(-M)}. \tag{23}$$

For Macquarie Island toothfish maturity is defined via length, in terms of a logistic rela-

tionship with $l_{50} = 1.396$ and $l_{95} = 1.858$ (Day *et al.*, 2015). As for weight-at-length the usual allometric relationship, $w_l = al^b$, is used where $a = 4.4 \times 10^{-6}$ and $b = 3.14$. To estimate maturity, m_a, and weight-at-age, w_a, for the population model we need to integrate over the distribution in length-at-age, $\Pr(l \mid a)$:

$$m_a = \int \Big[m_l * \Pr(l \mid a) \Big] \mathrm{d}l, \tag{24}$$

$$w_a = \int \Big[w_l * \Pr(l \mid a) \Big] \mathrm{d}l. \tag{25}$$

Once we have computed these age-based vectors we calculate the SSB-per-unit-recruit as follows:

$$SPR_{F=0} = \sum_a \tilde{n}_a w_a m_a. \tag{26}$$

For each MCMC sample from the (female) conditional age-at-length growth model we calculated the quantity in Eqn.(26). Posterior median (and 95% credible intervals) were 3.84 (3.26–4.46) (in units of tonnes $\times 10^{-6}$ per recruit); the posterior mean is the same as the median with a CV of 0.08. This is not a large amount of variability, but it does propagate through the assessment over time, and into the SSB from the very start, so it is not irrelevant - especially when uncertainty interacts with the harvest control rule as it does in this case.

One of the near-universal features of the von Bertalanffy growth model is that k and L_∞ are *strongly* negatively correlated (-0.96 in this case). If we were to fix one of the two parameters, k for example, what would be the reduction in uncertainty in key assessment variables? For the SPR example, this is simple and we fixed k at the posterior mode and re-ran the full MCMC routine. The posterior median (and 95% credible interval) for the unfished SPR was now 3.79 (3.52–4.08), with the same mean as the median and a CV of 0.04.

Perhaps surprisingly, the reduction in the uncertainty in unfished SPR alone was 50% in terms of the CV, just by fixing k - even given very strong negative correlation with L_∞. So it would seem that there will be a significant variance trade-off accompanying the fixing of even one of the growth parameters.

Discussion

The growth model is of fundamental importance to any length and age structured stock assessment. A number of approaches have been explored for Macquarie Island toothfish over the years, given the extensive set of age and length data from this fishery. The current default method is to estimate growth inside the assessment model, using the conditional age-at-length approach. For females, L_∞ has been fixed in the assessment given early problems

obtaining consistent and sensible estimates (Fay , 2010). The past two years assessments (Day *et al.*, 2014, 2015) have seen male estimates of k and L_∞ increasingly diverge from estimates obtained from external estimation frameworks which give consistent, precise and very similar answers.

As seen in previous work (Hillary *et al.*, 2014), the more complicated hierarchical length-at-age model (with an L_∞ for each animal) outperformed the basic one in terms of explaining the data, but the estimates were very similar. The conditional age-at-length model is, in many ways, more complicated and uses the length frequency and age-at-length frequency data together. We developed a detailed process error model for these data using an empirical Bayes approach to actively estimating the over-dispersion coefficients. While complex, the approach appeared to be validated when comparing the posterior predictive performance of the models with and without process error (over-dispersion). The model without process error consistently under-estimated the variability in both the length and age-at-length data. When comparing with the hierarchical length-at-age model, estimates of growth rate k were very similar; estimates of L_∞ were slightly lower for the age-at-length approach; estimates of t_0 were closer to zero for the age-at-length approach; and estimates of the stochastic variability in length-at-age (σ_p) were very similar as well. The variance in the parameter estimates were always greater for the age-at-length approach - particularly for both L_∞ and t_0. This is not necessarily a surprise given that ageing error is accounted for in the age-at-length approach, and will have a particular effect on the parameters related to where the data are most sparse: at the upper and lower ends of the age and length range.

In the assessment the fits to the actual length and conditional age-at-length data are acceptable. In one sense this is what really matters: how well are you fitting the ages and lengths in the actual data, not at the extremes. Additionally, given the very strong negative correlation between L_∞ and k the outcomes of the assessment and, crucially, the management advice are unlikely to be very strongly altered for fixed growth parameters. However, the estimates are clearly nonsensical given the life-history of the fish, and clearly at odds with both the externally obtained length-at-age and conditional age-at-length estimates. It is not the way the data are used that appears to be the problem, given the consistency of the external estimates, so that leaves only the interaction with other fixed parameters and additional data sets (and their weightings) in the assessment.

Overall, the length-at-age and conditional age-at-length approaches appear to give generally consistent and precise estimates of the growth parameters of interest. This is obviously comforting but it also gives us a crucial linkage between estimates of length-at-age from the "classical" approach and those obtained via the conditional length-at-age approach in the assessment. Prior to this work, differences could arise from the difference in approach

in the assessment, from assumptions made about key life-history parameters or processes (selectivity) within the assessment itself, or from other data sources within the assessment model. We have seen in this work that it if differences do arise, they do not appear to be driven by the statistical differences between the length-at-age and conditional age-at-length approaches. This leaves either model and parameter assumptions made within the assessment or the other data sources as the remaining drivers of different estimates. This has always been the major potential weakness of estimating the growth from within the assessment in the integrated framework. From analyses of the components of the objective function across data sets and values of male L_∞ the key data sets causing the issue are the tag total recaptures and spatial composition, with age-specific movement the most likely structural issue in the model (movement is assumed age-independent). The counter argument is that growth uncertainty *does* matter - especially when maturity is a function of length, as maturity-at-age can then be a source of "stealth" uncertainty. Even with accurate growth estimates such as these, the underlying CV in SSB-per-unit-recruit *just* from the uncertainty in weights and maturity-at-age (i.e. fixed M) is almost 10%. So we would, ideally, want to account for growth uncertainty within the assessment. Here, we seem to be hitting the ever-present bias/variance trade-off by trying to do so.

The current decision by the resource assessment group responsible for this stock is that in future we should fix both male and female asymptotic length in the assessment. While the cause of the growth parameter issue appears to be age-specific movement, caution must be exercised in further complicating the model in an attempt to fix the issue at present. Moving to age-specific movement, even a "simple" parametric approach like a logistic curve, means at least two more parameters and will require the length range of current recaptures to be far wider than it is at present (Day *et al.*, 2015). In time, with sufficient recaptures with longer times-at-liberty, the estimation of age-specific movement rates may become feasible. However, given the proposed solution results in performance diagnostics and general visual fits to the data that are close to indistinguishable from the case where male asymptotic length is estimated, the simplest solution should be the first choice.

References

Brooks, S.P., and Roberts, G.O. 1998. Assessing convergence of Markov chain Monte Carlo algorithms. Statist. and Comput. **8**: 319–335.

Candy, S.G., Nowara, G. B., Welsford, D. C., and McKinlay, J. P. 2012. Estimating an ageing error matrix for Patagonian toothfish (*Dissostichus eleginoides*) otoliths using between-reader integer errors, readability scores, and continuation ratio models. Fish. Res. **115**: 14–23

Casella, G. 1985. An introduction to Empirical Bayes data analysis. Amer. Statistician. **39**: 83–87

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. 2006. Deviance information criteria for missing data models. Bayesian Anal. **4**: 651–674.

Day, J., Haddon, M., and Hillary, R.M. 2014. Stock Assessment of the Macquarie Island fishery for Patagonian toothfish (*Dissostichus eleginoides*) using data up to and including August 2014. SARAG.

Day, J., Haddon, M., and Hillary, R.M. 2015. Stock Assessment of the Macquarie Island fishery for Patagonian toothfish (*Dissostichus eleginoides*) using data up to and including August 2015. SARAG.

Fay, G. 2011. Stock assessment of the Macquarie Island fishery for Patagonian toothfish (*Dissostichus eleginoides* using data up to and including June 2010.

Francis, R.I.C.C. 2011. Data weighting in stock assessment models. Can. J. Fish. Aquat. Sci. **68**: 1124–1138.

Gelman, A., Carlin, B.J., Stern, H.S., and Rubin, R.B. 2006. Bayesian data analysis. Chapman and Hall, London, UK.

Hillary, R. M., Wayte, S., Day, J., and Haddon, M. 2014. Appropriate growth models for Macquarie Island toothfish. SARAG.

Meng., X. 1994. Posterior predictive p-values. Ann. Stat. **22**: 1143–1160.

Methot., R.D., and Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fish. Res. **142**: 86–99.

Morton, R., and Bravington, M. V. 2008. Comparison of methods for estimating age composition with application to Southern Bluefin Tuna (*Thunnus maccoyii*). Fish. Res. **93**: 22–28.

Neal, R. M. 2011. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (eds Brooks, Gelman, Jones, Meng). Chapman and Hall/CRC, London, UK.

Piner, K. R., Lee, H.-H., and Maunder, M. N. 2015. Evaluation of using random-at-length observations and an equilibrium approximation of the population age structure in fitting the von Bertalanffy growth function. Fish. Res. `http://dx.doi.org/10.1016/j.fishres.2015.05.024`.

Appendix

For the conditional age-at-length model we implemented a variant of one of the suite of so-called Hamiltonian MCMC (HMC) algorithms described in (Neal, 2011). The HMC general approach is very different to either of the Metropolis-Hastings or Gibbs sampling (and mixtures thereof) approaches used in the length-at-age models. The first step is to consider the distribution of interest in terms of what in statistical mechanics is referred to as a *canonical distribution* of a random state variable x, defined via a suitable energy function, $E(x)$. The canonical distribution (over states x) has the following probability distribution:

$$P(x) = \frac{1}{\Xi} \exp\left(-\frac{E(x)}{T}\right),$$

where Ξ is a normalisation constant so $P(x)$ integrates to 1 over the support of x, and T is the *temperature* of the system. In simulated annealing T plays a central role in the optimisation algorithm, but here we simply set it to $T = 1$ for all cases. It is trivial to work instead with a probability distribution (via Bayes' theorem combining the likelihood and prior) by setting $E(x) = -\log \pi(x) - \log \Xi$ so viewing the distribution of interest, the posterior distribution $\pi(\bullet)$, in this way is no barrier to what follows.

The Hamiltonian defines a joint energy distribution for "position" q and "momentum" p variables:

$$P(q, p) = \frac{1}{\Xi} \exp\left(-\frac{H(p, q)}{T}\right),$$

and choosing an additive form for the Hamiltonian, where $H(p, q) = U(q) + K(p)$, means the joint density is defined as follows:

$$P(q, p) = \frac{1}{\Xi} \exp\left(-\frac{U(q)}{T}\right) \exp\left(-\frac{K(p)}{T}\right),$$

so q and p are independent (with their own energy functions and associated canonical distributions). The key part of the process is to consider q to represent the parameters of the

posterior distribution of interest, and p to simply be dummy variables that are useful only in the sense that they mean the joint dynamics of the position and momentum variables are Hamiltonian. So, with $T = 1$, we set the "position" energy function to be

$$U(q) = -\log[L(q \mid Y)p(q)],$$

where $L(q \mid Y)$ is the likelihood given the data Y, and $p(q)$ is the prior for the estimated parameters.

HMC samples from the (canonical) distribution of both p and q and, while we are uninterested in the p "momentum" variables, we do need to define their distribution also. The currently accepted best practice here is a quadratic energy function (effectively a Gaussian distribution):

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i},$$

where d is the number of parameters in the model of interest and the m_i variance parameters play a role in the effectiveness of the HMC algorithm (Neal, 2011).

At a general level, the HMC algorithm is a two-step process. The first step is to randomly draw new momentum variables (the p) from their Gaussian distribution independently of the current position variables (the q which we care about). Since q has not changed, and the p are drawn from their correct marginal distributions (which are independent of the q by definition) which leaves the (canonical) joint distribution invariant.

The second step uses a Metropolis proposal algorithm to update the position variables q. Given the current state (q, p), we propose new values of the two variables (q', p') and accept this new proposal according to the usual acceptance probability:

$$\min[1, \exp(-H(q', p') + H(q, p))] = \min[1, \exp(-U(q') + U(q) - K(p') + K(p))].$$

The more complicated part of this process, relative to say the random-walk Metropolis algorithm (Gelman *et al.*, 2006), is how to generate the new joint proposal (q', p'). The invariance of the trajectories on the (volume-preserving) Hamltonian surface is what makes the process work, so the proposal generator must also maintain this invariance. Hamiltonian dynamics are defined as follows:

$$\begin{aligned}\frac{dq}{dt} &= \frac{\partial H}{\partial p},\\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q},\end{aligned}$$

where volume is preserved (because $\nabla \cdot (\dot{q}, \dot{p}) = 0$). To generate new proposals we need to integrate the Hamiltonian dynamics of (q, p) forward in such a way as to be volume preserving. The simplest way to do this is via a modification of the Euler method, called the *leapfrog* method (Neal, 2011).

We choose a suitable time step-size, ϵ, and the scheme to advance $(q(t), p(t))$ to $(q(t + \epsilon), p(t + \epsilon))$ proceeds as follows:

$$\begin{aligned}p_i(t + \epsilon/2) &= p_i(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t)),\\ q_i(t + \epsilon) &= q_i(t) + \epsilon\frac{p_i(t + \epsilon/2)}{m_i},\\ p_i(t + \epsilon) &= p_i(t + \epsilon/2) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}(q(t + \epsilon)),\end{aligned}$$

and i denotes the particular element of the position/momentum vectors. The leapfrog algorithm is both volume-preserving (and reversible) making it the most parsimonious way to simulate the Hamiltonian dynamics correctly. To generate new proposals (q', p') given our current state (q, p) we simply take $\mathcal{L}$ steps of step-size ϵ (to a total "time" of $\epsilon\mathcal{L}$) via the leapfrog algorithm. One note of caution is that the leapfrog algorithm requires the first derivatives of the posterior distribution of interest to be used. Given we also need the accuracy of the algorithm to sensibly move around the Hamiltonian surface, using finite differences to calculate these derivatives is not advised, and almost never will we have analytical expressions for them either. In this particular case we use automatic (or algorithmic) differentiation to calculate these first derivatives to machine precision. The `CppAD` library `http://www.coin-or.org/CppAD/`, a `C++` based open-source package, is used in this instance.

Performance (and related tuning) of the HMC algorithm is most strongly linked both the step-size, ϵ, and the number of steps to take, $\mathcal{L}$, for a given parameter update. There is some linkage to the m_i too, but this can be mitigated with sensible scaling frames for the estimated parameters. It ϵ is too high, relative to the variance of the key parameters in question, the integration scheme becomes unstable. If ϵ is too low then we are needlessly adding

computational time to the algorithm. The number of steps $\mathcal{L}$ needs to be large enough so that the algorithm moves around the distribution and mixes well, attaining good acceptance rates. Equally, too large and it could, eventually, begin retracing the steps or even simply be staying effectively in the same region of state-space. The key with tuning the algorithm is experience and experimentation, more so than the random-walk Metropolis algorithm. That being said, the recommendations for tuning the algorithm are well established (Neal, 2011) and, at least from experience, are not unduly difficult to implement.

For the specific conditional age-at-length problem every parameter bar t_0 was log-transformed as continuous unbounded variables are required for the HMC algorithm. The log transformation has the additional benefit of making it fairly simple to choose $m_i \equiv 1$ for the momentum variable variance and tune only ϵ and $\mathcal{L}$. For the step-size we chose a uniform distribution: $\epsilon \sim U[0.0015, 0.0025]$. We take a similar approach with the number of leapfrog steps to take: $\mathcal{L} \sim U[200, 300]$. These two parameters are resampled after each set of $\mathcal{L}$ steps has been taken, not every single update from t to $t + \epsilon$. This randomisation of the HMC performance parameters is recommended to avoid the stability and mixing pitfalls outlined previously.

The over-dispersion parameters are estimated first, using the empirical Bayes approach outlined in the main text. The marginal (negative) log-likelihood (and its derivatives calculated via automatic differention) are used in a `C++` based BFGS minimisation routine and the over-dispersion coefficients are estimated herein via marginal maximum likelihood estimation. The best estimates of these over-dispersion parameters are then used (and assumed fixed) when the HMC algorithm is implemented to sample from the joint posterior of the growth and prior age distribution parameters. The algorithm was very effecient and achieved excellent acceptance rates (80–85%), with good mixing characteristics and little to no autocorrelation. The full 10,000 simulations arising from the HMC algorithm was thinned equally by a factor of 10, leaving 1,000 samples that were deemed converged to the posterior via established methods (Brooks and Roberts, 1998). One full run of the HMC algorithm took around 3–4 minutes on a laptop running four 2.6 GHz processors.

Tables

Parameter	L_∞	k	t_0	σ_p
		Basic length-at-age		
Males	1.27 (0.03)	0.07 (0.07)	-2.28 (0.1)	0.14 (0.02)
Females	1.69 (0.04)	0.05 (0.08)	-2.52 (0.09)	0.15 (0.02)
		Hierarchical length-at-age		
Males	1.31 (0.15)	0.066 (0.07)	-2.53 (0.09)	N/A
Females	1.81 (0.16)	0.046 (0.07)	-2.55 (0.08)	N/A
		Conditional age-at-length		
Males	1.21 (0.06)	0.08 (0.1)	-1.23 (0.18)	0.15 (0.04)
Females	1.7 (0.07)	0.05 (0.09)	-1.5 (0.15)	0.15 (0.04)

Table 1: *Posterior mean (and CV in brackets) estimates for the male and female growth curve parameters and for all three model frameworks explored.*

L_∞ (male)	Total	Length	Age	Tag (comp.)	Tag (recap.)	Recruitment
1.3	7.1	0.0	0.7	5.1	10.0	0.2
1.65	3.8	2.2	0.1	3.3	7.1	0.2
2	2.4	3.6	0.0	2.3	5.3	0.1
$18.51^\star$	0.0	7.6	1.4	0.0	0.0	0.0

Table 2: *Objective function breakdown for each of the data sources and across the four male L_∞ scenarios explored in the latest assessment (with the $\star$ designating the actual estimate). For each objective function element, and the total, the minimum value is subtracted from all the others, so the interpretation is positive values are indicative of a worse fit and zero being the best.*

Figures

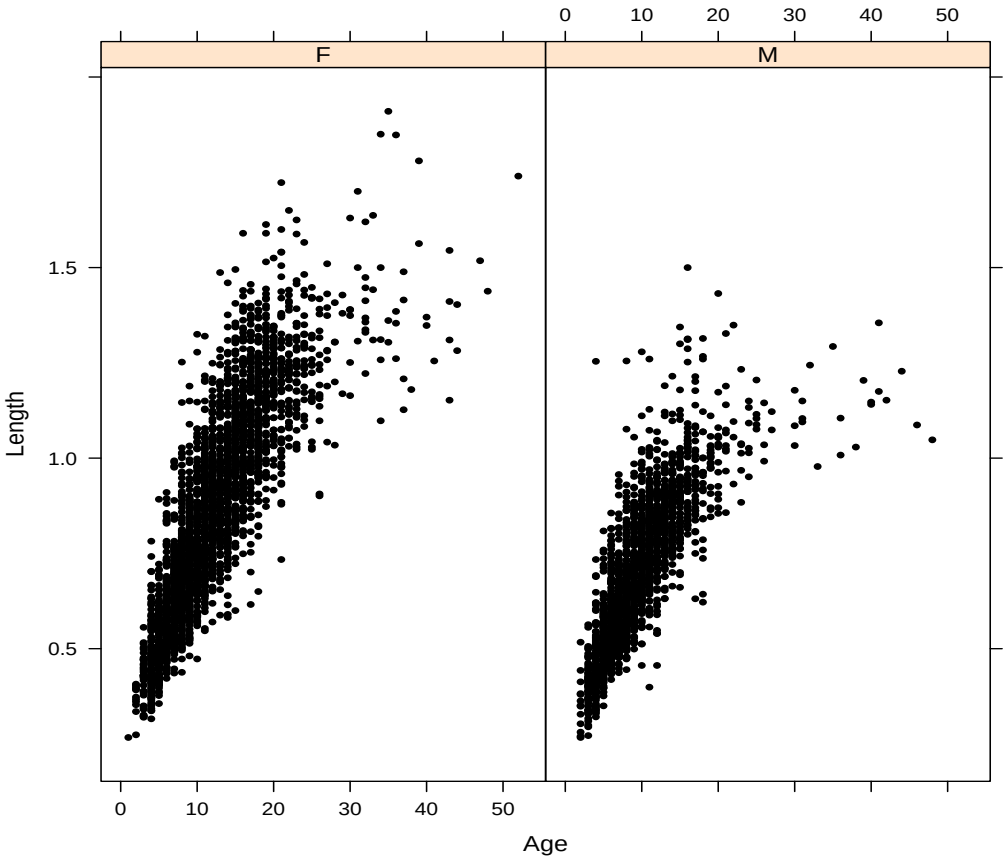

Figure 1: *Female (left; N=1,921) and male (right; N=2,756) length-at-age measurements for Macquarie Island toothfish from 1996 to 2015.*

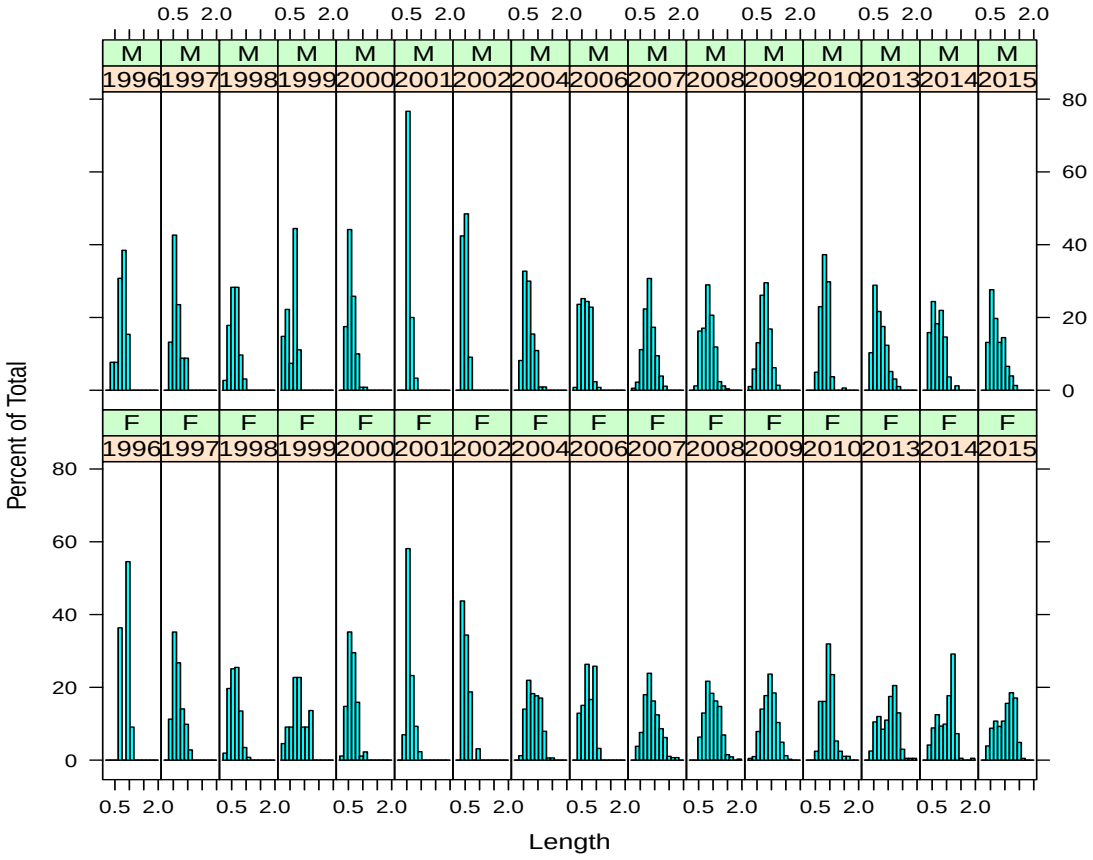

Figure 2: *Length frequency of age samples (by sex).*

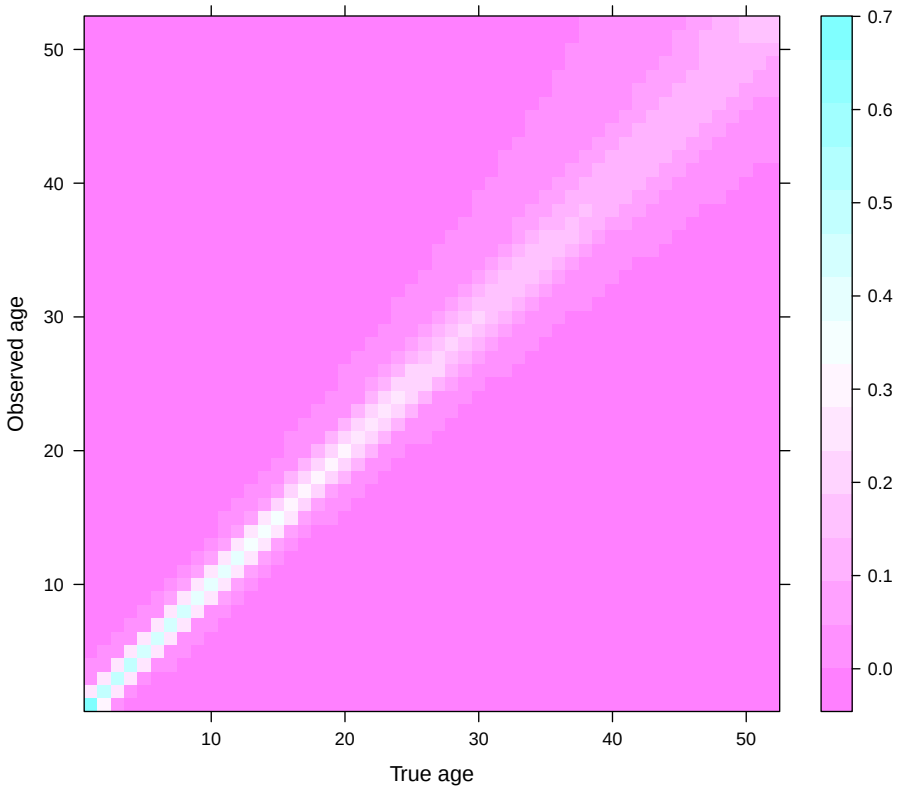

Figure 3: *Ageing error matrix used for Macquarie Island toothfish. The x-axis denotes the true age, and the y-axis the observed age*

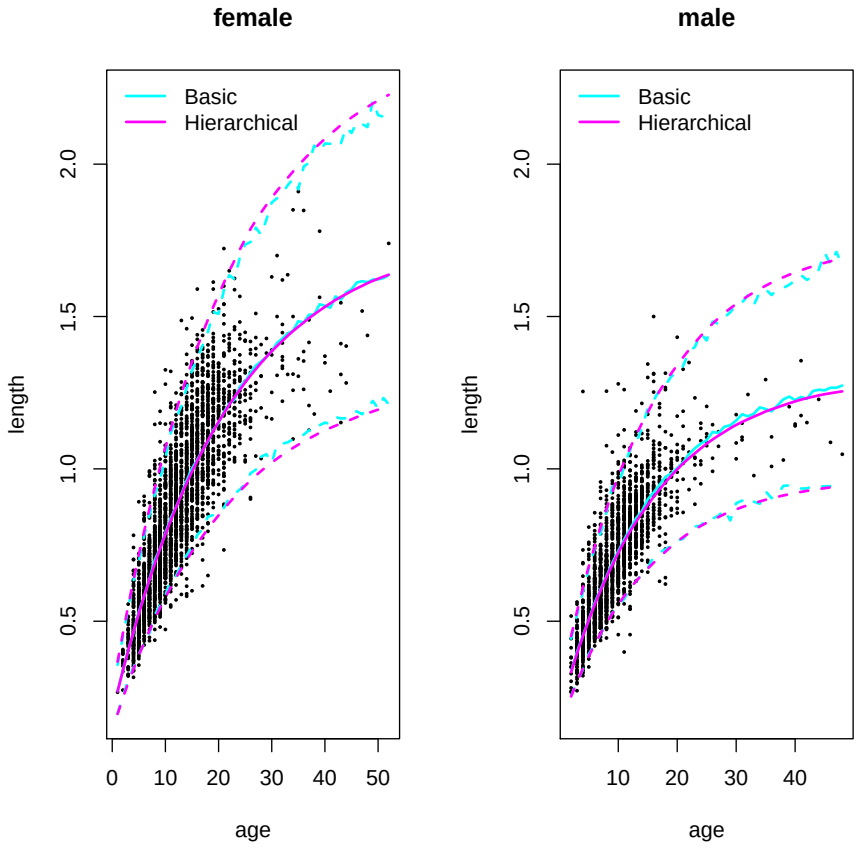

Figure 4: *Posterior predictive median (full line) and 95% credible interval (dashed lines) for the basic and hierarchical length-at-age model as applied to the female (left) and male (right) data (circles).*

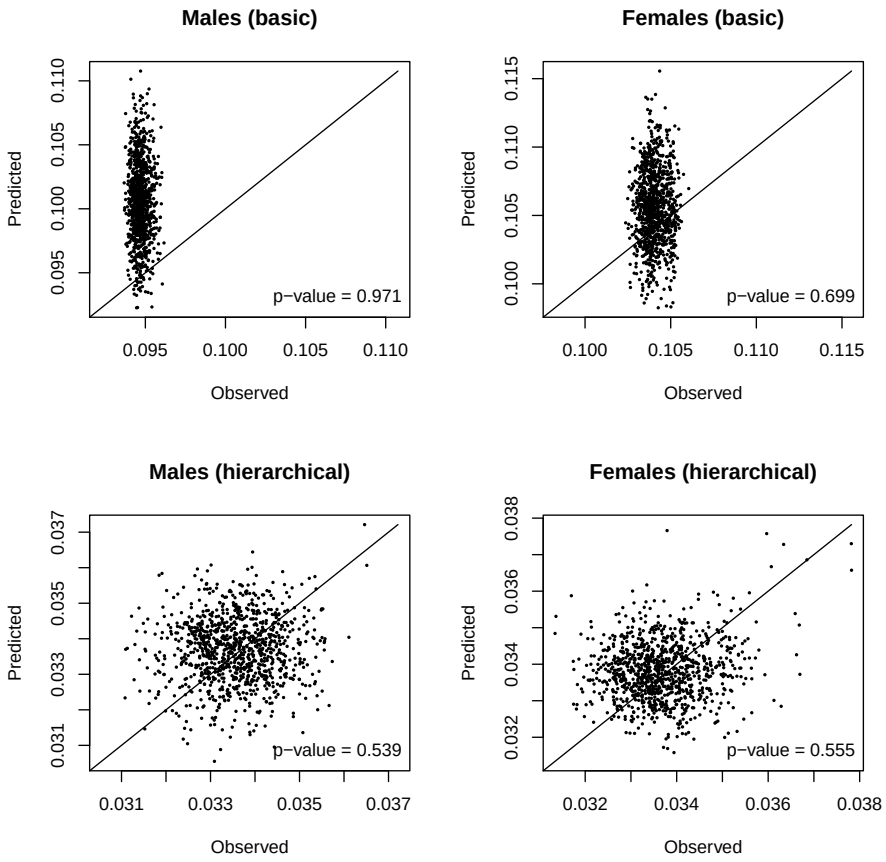

Figure 5: *Posterior predictive analysis for the basic (top) and hierarchical (bottom) length-at-age model as applied to the female (left) and male (right) data.*

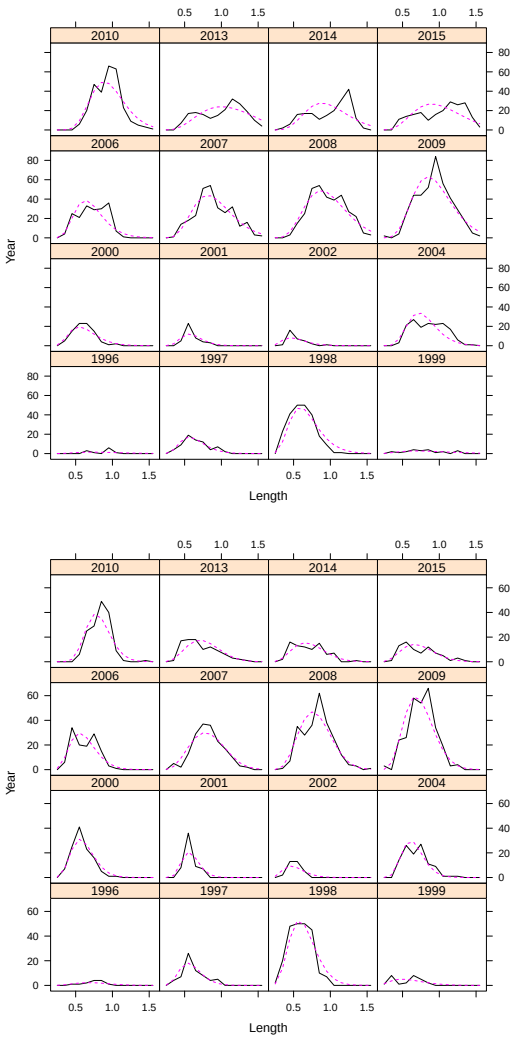

Figure 6: *Observed (black line) and estimated (dotted magenta line) length data for females (top) and males (bottom) at the posterior median parameter estimates.*

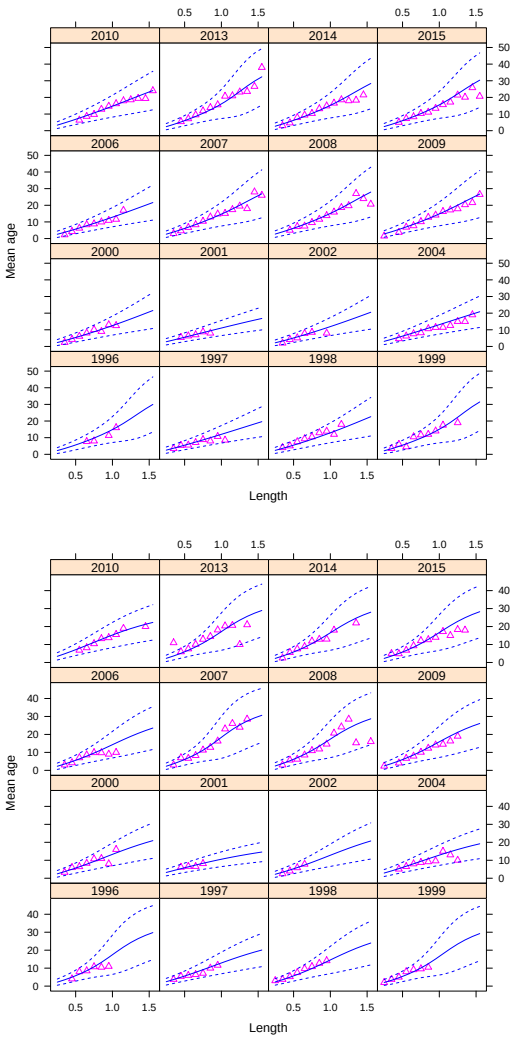

Figure 7: *Observed (magenta triangle) and posterior median (blue line) and 95% credible interval (dotted blue line) mean age for each length bin, for females (top) and males (bottom).*

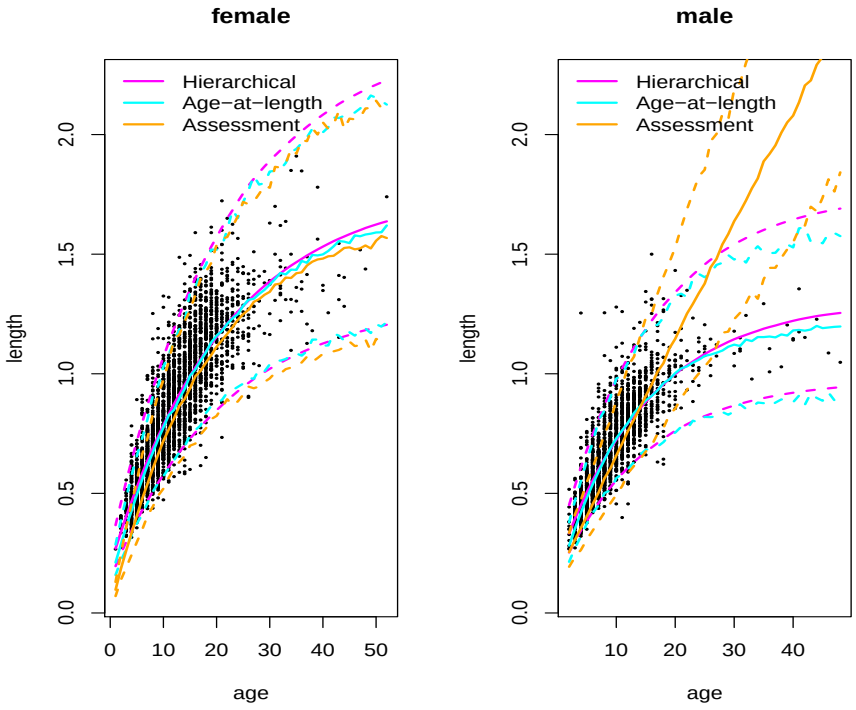

Figure 8: *Predictive intervals for the paired length-age data for the hierarchical length-at-age, conditional age-at-length, and current stock assessment growth curves.*