


Overcoming long Bayesian run times in integrated fisheries stock assessments

Cole C. Monnahan ^{1,2,*}, Trevor A. Branch¹, James T. Thorson³, Ian J. Stewart⁴, and Cody S. Szuwalski⁵

¹*School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98105, USA*

²*Departamento de Oceanografía, Universidad de Concepción, Casilla 160-C, Concepción, Chile*

³*Alaska Fisheries Science Center, National Marine Fisheries Service—NOAA, Seattle, WA 98115, USA*

⁴*International Pacific Halibut Commission, 2320 West Commodore Way, Suite 300, Seattle, WA 98199-1287, USA*

⁵*Marine Science Institute and Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 93105, USA*

*Corresponding author: tel: 1-360-303-2428; e-mail: monnahc@uw.edu.

Monnahan, C. C., Branch, T. A., Thorson, J. T., Stewart, I. J., and Szuwalski, C. S. Overcoming long Bayesian run times in integrated fisheries stock assessments. – ICES Journal of Marine Science, doi:10.1093/icesjms/fsz059.

Received 22 October 2018; revised 27 February 2019; accepted 11 March 2019.

Bayesian inference is an appealing alternative to maximum likelihood estimation, but estimation can be prohibitively long for integrated fisheries stock assessments. Here, we investigated potential causes of long run times including high dimensionality, complex model structure, and inefficient Bayesian algorithms for four US assessments written in AD Model Builder (ADMB), both custom built and Stock Synthesis models. The biggest culprit for long run times was overparameterization and they were reduced from months to days by adding priors and turning off estimation for poorly-informed parameters (i.e. regularization), especially for selectivity parameters. Thus, regularization is a necessary step in converting assessments from frequentist to Bayesian frameworks. We also tested the usefulness of the no-U-turn sampler (NUTS), a Bayesian algorithm recently added to ADMB, and the R package *adnuts* that allows for easy implementation of NUTS and parallel computation. These additions further reduced run times and better sampled posterior distributions than existing Bayesian algorithms in ADMB, and for both of these reasons we recommend using NUTS for inference. Between regularization, a faster algorithm, and parallel computation, we expect models to run 50–50 000 times faster for most current stock assessment models, opening the door to routine usage of Bayesian methods for management of fish stocks.

Keywords: AD Model Builder, Bayesian inference, fisheries stock assessment, no-U-turn sampler, Stock Synthesis

Introduction

Fisheries stock assessment models are population dynamics models used to explore the effects of management actions on fish populations (Hilborn and Walters, 1992). Varied data availability, fish life histories, and fishery properties drive a wide range of models for fish dynamics. Currently, the state-of-the-art approach for data-rich stocks is integrated analysis, where non-linear statistical models incorporate multiple data sources (e.g. Maunder and Punt, 2013). Integrated models vary in dimensionality and internal structure (“complexity”), and our focus here is on the most complex

integrated stock assessments used for management. Integrated models are typically written in the programming framework AD Model Builder (ADMB; Fournier *et al.*, 2012), and can be purpose-built for a particular stock (e.g. Szuwalski and Turnock, 2016) or be generic such as the widely-used Stock Synthesis (Methot and Wetzel, 2013). Regardless of the specifics, integrated models attempt to infer biological and fishery processes from complex, varied data, and are a vital tool to inform management.

Integrated models have been estimated in both the frequentist and Bayesian statistical paradigms (Punt and Hilborn, 1997;

Maunder, 2003). There are key philosophical differences between these paradigms, such as the definition of probability, sources of information, and interpretation of uncertainty statements (e.g. de Valpine, 2009; Gelman et al., 2014), and practical difficulties with Bayesian integration (e.g. Thorson and Cope, 2017). Here, we focus on technical differences in the estimation of parameters and uncertainty. The frequentist approach is popular for complex models because it is generally faster and has reliable parameter estimates. However, Magnusson et al. (2013) showed that Bayesian estimates of uncertainty are more reliable than frequentist estimates and recommend Bayesian methods as the default, although Stewart et al. (2013) argued that both methods have advantages. The Bayesian approach also offers a formal way to incorporate prior information from previous studies or expert opinion, and is a natural framework for estimating probabilities of hypotheses and performing decision analyses (Punt and Hilborn, 1997). Despite this, prohibitively long run times of the order of days to months have limited their practical use for management of data-rich stocks (but see Grandin et al., 2016) because exploring model sensitivity and evaluating different cases during development or the review process is difficult (Cotter et al., 2004). Therefore, although there are compelling reasons to perform Bayesian analyses on stock assessments, the time needed to make inference is a major obstacle.

Differences in run times between the two paradigms are related to how inference is made. The frequentist approach uses only information from the data likelihood and involves estimating parameters using maximum likelihood. Uncertainty in parameters is estimated by assuming asymptotic normality and inverting the Hessian matrix evaluated at the maximum likelihood estimate (MLE). The uncertainty of derived quantities such as maximum sustainable yield is estimated using the delta method (Magnusson et al., 2013). For Bayesian inference, the likelihood is combined with prior distributions to form the posterior probability distribution. Posterior probability statements for parameters and derived quantities are then approximated from posterior samples, typically generated with Markov chain Monte Carlo (MCMC) algorithms (Gelman et al., 2014). Long Bayesian run times generally arise from the need to approximate integrals of a complex, high-dimensional probability distribution by evaluating the model hundreds of thousands to tens of millions of times via MCMC. The easiest way to reduce run time is to decrease time per evaluation through strategies such as buying faster computers or using courser approximations to the population dynamics (Monnahan et al., 2016; Szuwalski, 2016). Although helpful, these options are not likely to reduce run time by orders of magnitude as required. Instead, we focus on improving MCMC algorithm efficiency, which is a more general and promising approach.

The most popular algorithm within ADMB is a modified “random-walk” Metropolis algorithm (RWM; Metropolis et al., 1953), which is inefficient for higher dimensions and for hierarchical models. Recently, a new MCMC algorithm was added to ADMB (Monnahan and Kristensen, 2018) called the no-U-turn sampler (NUTS; Hoffman and Gelman, 2014), which efficiently samples from high-dimensional and complex posterior geometries and is widely used in diverse applied statistical fields (Monnahan et al., 2017). NUTS is a variant of the Hamiltonian Monte Carlo family of MCMC algorithms (HMC; Neal, 2011), which automates tuning of the step size and trajectory lengths by the analyst and this flexibility makes it a good option for generic inference. HMC algorithms, including NUTS, are powerful

because they use posterior gradient information to generate distant proposals and reduce autocorrelation. Importantly, NUTS is also more robust to bias arising from approximating the posterior distribution with finite MCMC samples because divergences warn when extreme curvature exists in the posterior that can lead to bias (Betancourt, 2017; Monnahan et al., 2017). A thorough explanation of NUTS is beyond the scope of this study, so we refer to interested readers to introductory material referenced above. The other HMC algorithms available in ADMB are generally less efficient and not explored here. Both RWM and NUTS are tuned to increase efficiency, and this is a key approach for reducing run time. The most important tuning parameter in ADMB is the use of information about the global geometry of the posterior (Supplementary Appendix A). In the HMC literature, this information is encoded in the “mass matrix”, and rotates and scales the posterior to improve sampling (Neal, 2011). A similar approach is used by ADMB for the RWM algorithm, which uses the estimated covariance matrix (i.e. the inverted Hessian matrix) as the mass matrix by default. In general, both algorithms will be more efficient when the posterior resembles independent standard normal distributions, and this will be true if the mass matrix closely approximates the posterior and is multivariate normal. If the mass matrix does not approximate the posterior well (we refer to this as “mismatched mass matrix”) then sampling will be less efficient (Supplementary Appendix A).

Efficiency can also be improved by changing the geometry of the posterior. One way to accomplish this is to reparameterize a component of the model, such as somatic growth, to a form more suitable for statistical estimation (e.g. Schnute, 1981). Another is to add informative priors or fix parameters (i.e. assuming a potentially estimable parameter to be constant), which we refer to as “regularization”, adding more information and effectively constraining the geometry of the posterior. Reparameterization and regularization influence run time because it effects the estimated mass matrix (Supplementary Appendix A), which is a key tuning parameter, but also because Bayesian inference requires integrating the entire posterior, and the geometry of areas with low posterior probability (the “tails” of the distribution) often determines efficiency. Although not explored in the stock assessment literature, geometric properties can have profound impacts on run time and may be undetectable by examining only the mode. Regularization is particularly important when stock assessments lack explicit priors, such as when designed for maximum likelihood estimation, and when models are overparameterized (i.e. have poorly-informed parameters) or are poorly-parameterized (correlations, which make estimation difficult).

Overparameterization can occur in many ways, especially given the diversity of stock assessments, but here we highlight two sources. First is fisheries selectivity, which and often has either the logistic form or a “dome-shaped” increase with a subsequent decrease (e.g. Sampson et al., 2011). Consequently, flexible parametric shapes such as the six-parameter “double-normal” shape are widely used (Methot, 2015). However, the way the curve is parameterized, and the defaults used by most analysts, present some challenges in the context of Bayesian integration. For instance, if the estimated curve is asymptotic, the parameter controlling the top of the descending limb has virtually no effect on the selectivity curve, leading to fat tails that do not negatively affect MLEs (which are often estimated at a lower or upper bound) but present a difficult geometry for Bayesian algorithms. The

second case is when true biological or fishery processes vary with time, which can have important consequences for management quantities (e.g. Thorson *et al.*, 2015b; Stewart and Monnahan, 2017). A variety of approaches are used to model these processes in stock assessments (e.g. Methot and Wetzel, 2013), including the use of random effects, meaning that processes are drawn from a hyperdistribution (e.g. Thorson and Minto, 2015). A common example is annual recruitment deviations around the stock recruit curve, but time-varying processes like catchability and selectivity can also be modelled with this approach (e.g. Thorson and Wetzel, 2015; Stewart *et al.*, 2016). Use of random effects typically greatly increases the dimensionality of the model, and estimating hypervariances is difficult and rarely done in frequentist ADMB models (Thorson *et al.*, 2015a). Sophisticated parametrizations for selectivity and time variation are more likely, to closely, reflect the reality of the biology and fishery, but the added structural complexities can have important negative consequences for the statistical integration necessary for Bayesian inference.

Despite the advantages of Bayesian inference, no previous studies have addressed the problem of prohibitively long run times of Bayesian inference in integrated assessments, and as a result, the application of Bayesian inference for fisheries management is limited. In this study, we investigate causes for long run times for a set of stock assessments varying in size, complexity, and structure (age- and length-based). We first highlight the critical role of posterior geometry away from the mode in determining run time, present practical guidelines for diagnosing geometric properties that lead to slow mixing and show how to mitigate these issues to substantially reduce run time. Then, we contrast the newly available NUTS algorithm to the status quo RWM algorithm in their relative efficiencies and ability to avoid parameter bias. Finally, we compare the estimated uncertainties in key derived management quantities between frequentist and Bayesian paradigms for four case studies.

Methods

Case studies

We chose our case studies to reflect the fact that integrated stock assessments can vary widely in size, complexity, and structure (Tables 1 and 2). We chose age- and length-structured models, custom built and Stock Synthesis models, and models with time-

varying components. Some chosen models were expected to mix efficiently, whereas others were expected to be slow.

Hake model: the Pacific hake (*Merluccius productus*) assessment (Grandin *et al.*, 2016) uses Bayesian inference for management via the RWM algorithm and has been the subject of a past study comparing frequentist and Bayesian inference (Stewart *et al.*, 2013). This model converges successfully using runs of 12 million thinned every 10 000 samples. Individual model evaluations are rapid because the model uses an empirical weight-at-age approach that does not need to track length dynamics internally (e.g. Kuriyama *et al.*, 2016).

Halibut model: Pacific halibut (*Hippoglossus stenolepis*) management uses four models to assess the stock, and here we use the coastwide model based on a short time-series (Stewart and Martell, 2015; Stewart *et al.*, 2016). This model is parameterized to use empirical weight-at-age data, random walks for temporal variation in catchability and selectivity, and early recruitment deviations to allow for non-equilibrium in the initial age structure.

Canary model: the canary rockfish (*Sebastes pinniger*) assessment (Thorson and Wetzel, 2015) includes three areas and has random deviations relating the proportion of recruitment going to each area, which is implemented as an additive effect in multivariate-logit space (Methot, 2015). This model estimates 12 selectivity curves, many with the double-normal pattern, and has conditional age-at-length data that causes very slow model evaluations.

Snowcrab model: the model for Eastern Bering Sea snow crab (*Chionoecetes opilio*) is size-structured, was custom built for this particular stock (Szuwalski and Turnock, 2016), and has more parameters than any of the other models ($n = 334$). The model tracks numbers at size by sex, maturity state, and shell condition, and simultaneously estimates a size transition matrix based on laboratory observations of pre- and post-moult lengths.

Calculating baseline efficiency

The initial efficiency (defined below) of the case studies was estimated as a baseline to compare relative improvements. For each model, we calculated efficiency as if from a single RWM chain, initialized from posterior draws from a previous run (Supplementary Appendix B) for 3 million iterations, discarding

Table 1. Summary of case studies used.

Model name	No. of parameters	Speed ($s \ 1000^{-1}$ evals)	Brief description	Species and reference
Hake	217	8.71	MCMC results used for management, empirical weight-at-age, Stock Synthesis	Pacific hake; <i>Merluccius productus</i> (Grandin <i>et al.</i> , 2016)
Halibut	195	24.06	Time-varying catchability, empirical weight-at-age, Stock Synthesis	Pacific halibut; <i>Hippoglossus stenolepis</i> (Stewart <i>et al.</i> , 2016)
Canary	304	188.10	Time-varying growth, three areas with different exploitation history but no movement, natural mortality varies by age for males, complex selectivity with 31 fleets, Stock Synthesis	Canary rockfish; <i>Sebastes pinniger</i> (Thorson and Wetzel, 2015)
Snow crab	334	18.57	Length-structured, custom built, considerations for sex, maturity state, and shell condition, growth per moult data available	Eastern Bering Sea snow crab; <i>Chionoecetes opilio</i> (Szuwalski and Turnock, 2016)

Speed is how many seconds 1000 model evaluations take and is calculated as warmup and sampling time (but not optimization) divided by the total iterations during a RWM runs in which gradients are not calculated.

Table 2. The effect of regularization (adding priors and turning off estimation of parameters) on MLEs (and standard errors) of key management targets.

Model	Quantity	Original	Regularized	% Change
Hake	Depletion (2015)	0.71 (0.20)	NA	NA
	OFL (2015)	2.51E+6 (6.43E+5)	NA	NA
	MSY	8.41E+5 (2.50E+5)	NA	NA
Halibut	SSB (2000)	4.67E+5 (3.17E+4)	4.64E+5 (2.63E+4)	−0.6% (−17.1%)
	SSB (2010)	1.82E+5 (1.56E+4)	1.79E+5 (1.37E+4)	−1.6% (−12.0%)
	SSB (2015)	1.90E+5 (1.88E+4)	1.86E+5 (1.69E+4)	−2.0% (−10.1%)
Canary	Depletion (2015)	0.63 (0.08)	0.63 (0.08)	0.8% (−0.2%)
	OFL (2015)	1952.72 (293.43)	1969.56 (292.43)	0.9% (−0.3%)
	SSB (2015)	3297.16 (274.23)	3292.75 (271.59)	−0.1% (−1.0%)
Snowcrab	Depletion (2015)	1.38 (0.12)	1.41 (0.12)	1.8% (−5.5%)
	OFL (2015)	28.28 (3.50)	29.82 (3.51)	5.4% (0.3%)
	MSY	300.92 (16.68)	311.69 (16.64)	3.6% (−0.3%)

Spawning-stock biomass (SSB), depletion (biomass relative to unfishable state), overfishing limit (OFL), and MSY are common management metrics on the US West Coast.

the initial 25% as a burn-in period during which the algorithm tunes and the samples are not valid, then thinning to save every 1000th iteration, which helps post-processing but discards information (Link and Eaton, 2012). We used the R package `adnuts` for these analyses, which provides a convenient framework and improved workflow for Bayesian inference in ADMB within the R software framework (Monnahan, 2018; R Core Team, 2018). We estimated the number of effective samples for each model parameter, which accounts for autocorrelation, using the function `monitor` in the `rstan` package (Stan Development Team, 2018) then defined MCMC efficiency as the minimum effective sample size of post-warmup samples (across parameters) divided by total run time. This can roughly be thought of as the time to obtain an independent sample from the posterior, and is a standard approach used in other studies (e.g. Hoffman and Gelman, 2014; Monnahan et al., 2017). We excluded compilation and optimization time but included warmup and sampling periods. Models were also compared using the time required to obtain 1000 effective samples, which is usually large enough to make good approximations for many key management quantities, including relative probabilities in the tails of the posterior with little Monte Carlo error. We provide R code demonstrating how to calculate both the efficiency and time to get 1000 samples in our demo workflow (Supplementary Appendix B). As with any Bayesian analysis, before making management inference in a real analysis, the samples should also be checked for signs of non-convergence using standard diagnostics such as potential scale reduction \hat{R} (Gelman et al., 2014). We ran parallel chains with `adnuts`, but calculated baseline efficiency as if from a single chain to simulate the traditional approach of using a single command line run to obtain posterior samples.

Improving efficiency with regularization

The first step in diagnosing inefficiencies was to examine the baseline “pilot” chains, using the workflow described above, and visually assessing geometric issues with the posterior by plotting pairwise posterior correlations for the slowest mixing parameters. Based on this feedback, we then used the following guidelines to regularize the posterior. We fixed parameters (i.e. assume a constant value) at their MLE when standard errors were unreliably estimated (common at bounds), added stronger priors or fixed

parameters, which were not informative, or reparameterized where possible (e.g. if using double-normal selectivity curves, but selectivity curves are asymptotic, convert them to logistic curves, Supplementary Appendix B). After regularizing, we re-ran the pilot chains and iterated this regularization process, usually around five times, until each case study showed well-behaved geometries (i.e. no parameters with very low effective sample sizes). Clearly, it was beyond the scope of this paper to reformulate four stock assessment models while maintaining continuity with management practices. Instead, we strongly regularized the models with the goal of maintaining similar behaviour at the mode but improving it in the tails. These regularized models are a proof of concept for potential improvements and may differ somewhat from official stock assessments used for management.

Comparing performance between RWM and NUTS

We assessed efficiency for our regularized case studies by running RWM and NUTS chains with the default mass matrix (i.e. the estimated covariance at the MLE, Supplementary Appendix A). NUTS chains were run for 3000 iterations and no thinning (because the NUTS chains have low autocorrelation) and using a warmup of 20% (<50% recommended in Stan because mass matrix adaptation is not done, see Supplementary Appendix A). We chose RWM chain lengths to ensure runs were similar in time duration to NUTS chains, and we compared the efficiency of the algorithms. Lastly, we re-ran NUTS with an updated mass matrix calculated as the empirical covariance of posterior samples from a previous run, which easy to do with `adnuts` for both algorithms (Supplementary Appendix B). We expected that using an improved tuning parameter in this way would more efficiently sample from the posterior when the estimated covariance did not accurately reflect the posterior geometry (Supplementary Appendix A). This procedure does not affect the resulting posterior, only the efficiency at which samples are generated. Since initial exploration showed no improvement for the RWM algorithm, we only focused on improvements to NUTS.

Results

Calculating baseline efficiency

We found the RWM pilot chains of the original models and default settings mixed poorly and failed to converge ($\hat{R} > 1.1$) for

Table 3. The estimated time to get 1000 effective samples for different models and algorithms and speed relative to the original model version (parentheses).

	Hake	Halibut	Canary	Snowcrab
Original	18.6 h (1)	12.4 months (1)	187.5 months (1)	38.7 months (1)
Regularized	NA (NA)	0.7 days (507)	39.1 days (144)	68.1 days (17)
RWM default	4.5 h (4)	1.2 days (301)	14.3 days (394)	9.6 days (121)
NUTS default	2.1 h (9)	120 min (4466)	44.8 days (126)	2.8 days (421)
NUTS updated	15 min (74)	39 min (13 741)	12.5 h (10 772)	23 min (72 706)

This is extrapolated from the estimated effective samples per time. The “Original” and “Regularized” versions are the pilot chains using the default ADMB workflow of a single chain using the Metropolis algorithm (RWM). The last three rows are the regularized model versions and assume four parallel chains. The RWM and NUTS “Default” chains use the inverted Hessian for the mass matrix (a tuning parameter), whereas the NUTS “Updated” chain uses an estimated mass matrix from a previous run. Note that these are rough approximations because they are based on estimated effective sample size, which are highly variable.

all but the Hake model. These results cannot be used for inference but can be extrapolated to determine it would take at minimum days and up to more than 2 years for the Canary model to generate 1000 effective samples (Table 3). Long runtimes were partly because each iteration was slow, particularly for the Canary model, which uses conditional age-at-length data (Table 1). However, a more important component was the overparameterization that occurred in all but the Hake model (Figure 1). For instance, the double-normal selectivity in the Canary model had long tails (Figure 2). Another common occurrence was a mismatched mass matrix, sometimes caused by an MLE at a bound, resulting in the underestimation of the variance of that parameter, or a poorly-informed parameter for which the variance was vastly overestimated (Figure 3), resulting in a poorly tuned algorithm.

Improving efficiency with regularization

Another mismatched mass matrix issue we found was locally varying correlations between parameters, as typified by early recruitment deviations in the Halibut model (Figure 4), where the model could not distinguish age classes when the data supporting a large recruitment come from sparse age distributions. In such cases, a large recruitment in year y and a small one in year $y+1$ or a small one in y and large one in $y+1$ can explain equally plausibly the subsequent number of fish observed. Ageing error can also lead to difficulty in distinguishing recruitment strength. The core issue is not a single correlation but correlation between sequential parameters: y is correlated with $y+1$, $y+1$ to $y+2$, etc., which causes an extreme posterior geometry (Supplementary Figure S1) that is poorly approximated by a global mass matrix. In such cases, the RWM chains were biased because they were unable to explore part of the posterior (Figure 4). We added arbitrary normal priors to these recruitment deviations during regularization to eliminate this geometry (Figure 4), an approach similar to VPA models where it was sometimes assumed that the oldest age classes at the beginning of the modelled period were equal to the mean recruitment decayed to that age (e.g. Quinn and Deriso, 1999). We found similar correlations for some pairs of selectivity parameters, but these were not as detrimental to mixing as the recruitment parameter correlations in the Halibut model.

The regularization process detailed above took about five iterations to successfully update the models, and steps differed among the three models (Supplementary Appendix C). We constrained a variety of selectivity parameters and early recruitment deviations in the Halibut model, and selectivity and recruitment

apportionment deviations in the Canary model. The Snowcrab model required the greatest variety of changes, including growth parameters, mortality deviations, and recruitment deviations, but few changes to selectivity parameters. After following the regularization procedure (except for the Hake model), the models mixed substantially better (Figure 5), passing convergence tests and with relatively minor changes to management outputs (Table 2). The Halibut, Canary, and Snowcrab models were 507, 144, and 17 times faster after regularization compared with the status quo MCMC approach, respectively (Table 3), and had $\hat{R} < 1.1$ and high effective sample sizes.

Comparing performance between RWM and NUTS

NUTS was generally faster than RWM on the regularized models using default settings (i.e. using the inverted Hessian as mass matrix) and passed convergence tests ($R_{hat} < 1.1$) as well as minimal divergences and no max treedepths exceeded. When using an updated mass matrix, NUTS speed improved again over the default settings up to 172 times faster for the Snowcrab model (Table 3) and had $< 1\%$ divergences and treedepths well below the maximum of 12. Taken together, the improvements to the models from the best-case scenario (NUTS with updated mass matrix and four parallel chains) vs. the status quo (single RWM chain using inverted Hessian mass matrix) there were substantial improvements: 74, 13 741, 10 772, and 72 706 times faster for the four models, reducing run times to obtain 1000 independent samples from a range of 18.6 h—years to 15 min—12.5 h for the four models (Table 3). The updated NUTS chains generally mixed well with few iterations needed for all models. Only the Canary model required more than an hour for sufficient convergence under ideal circumstances because of the increased computational time from the conditional age-at-length structure (ten times longer than other models; Table 1).

Differences in key management quantities between statistical paradigms were small except for the Hake model (Figure 6), where Bayesian posterior medians were much higher (between 14.4 and 24.9%) while the corroborating Stewart *et al.* (2013). Halibut model was consistently higher by ~ 4 –5%. In contrast the differences in the Canary model were negative for depletion and the overfishing limit (-3.4 and -7.1%) but positive for MSY (1.1%). The Snowcrab model had the smallest differences, all less than $\pm 2\%$.

Discussion

Bayesian inference for data-rich stocks is rarely used for applied management advice because of prohibitive run times. Our goal

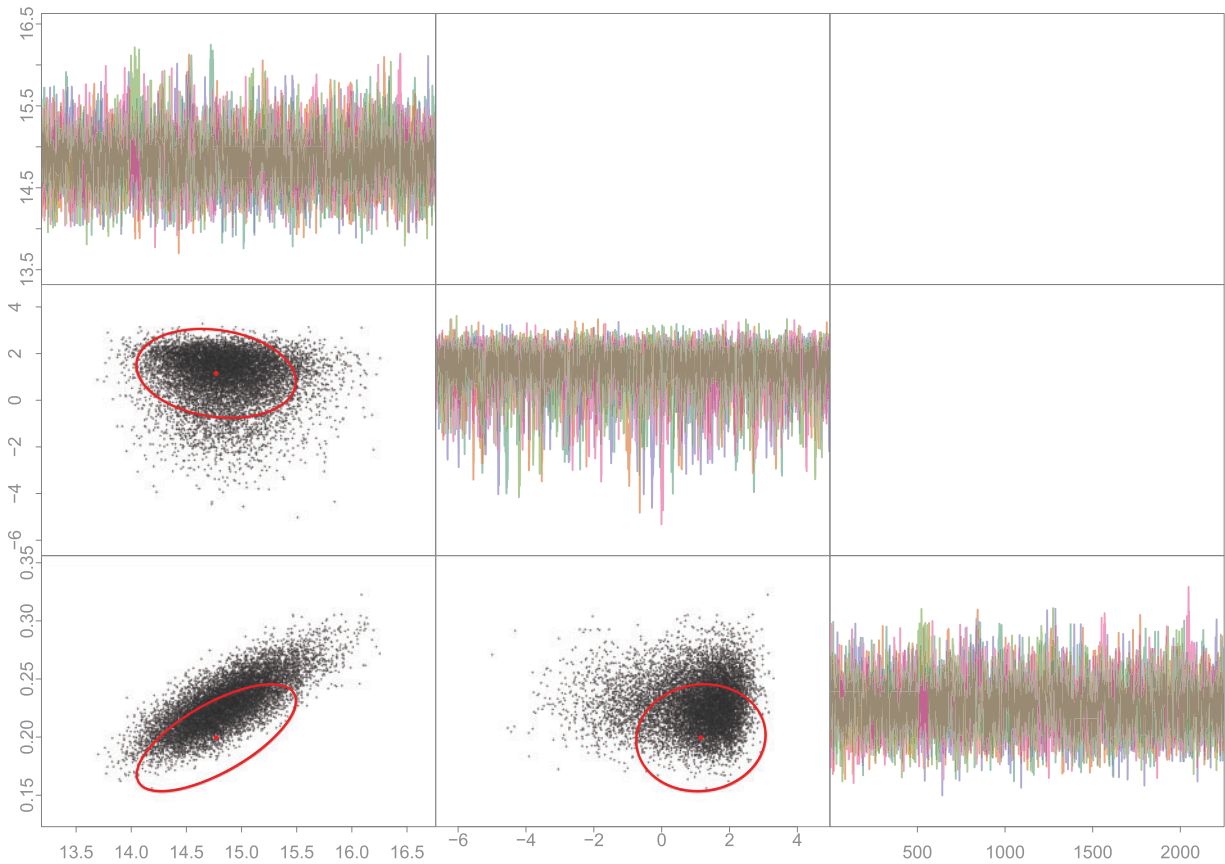


Figure 1. Diagnostic plots for the three slowest mixing parameters from the Hake model, comparing five RWM pilot chains to MLEs. The diagonals show traces of the five chains. The scatterplots show pairwise posterior samples (black dots) and bivariate 95% confidence regions (ellipses and single dots) from the inverted Hessian, which is used as the mass matrix, a key tuning parameter. The MCMC samples were thinned by 1000 and the chains initiated from the mode.

here was to identify causes for long run times, explore solutions, and quantify the potential reduction. We found the biggest cause is not model size or complexity, as often thought. Instead, overparameterization (i.e. poorly-informed parameters) resulted in difficult posterior geometries in the context of numerical integration. We demonstrated that regularization (i.e. constraining the model with priors or fixing parameters) coupled with well-tuned NUTS chains run in parallel can reduce run time from months to hours, suggesting that complex models with thousands of parameters could be viable if they are developed and parameterized for Bayesian inference. Certainly, taking most existing models “off-the-shelf” will require some work by an analyst for fast run times. Nevertheless, once this process is completed Bayesian inference is feasible for many models in a management and research framework.

There are important challenges to adapting NUTS for use in real-world stock assessments with a review process requiring fast inference for alternative model scenarios. In this context, the biggest barrier is how to develop models that are flexible but remain fast, which is a challenge because performance is sensitive to the estimated mass matrix. Exploring alternative configurations, which is relatively easy for Stock Synthesis models, may lead to mismatched mass matrices and require redoing the regularization process to a degree. More generally, regularization may be straightforward for parameters where biological priors can be specified, but for other parameters it is more challenging. For

instance, specifying a prior for unfished recruitment (R_0) is difficult, and the effect of a prior in log space may be quite different from in natural space (Punt and Hilborn, 1997; Thorson and Cope, 2017). Similarly, double-normal selectivity parameters typically have uniform priors, but the implied prior on selectivity itself is difficult to predict and should be explored (Supplementary Figure S2). Another example is the recruitment deviations in the Halibut model (Figure 4, Supplementary Figure S1), regularization of which had no clear technical solution within the Stock Synthesis framework. However, these difficulties are not specific to stock assessments and are a key part of ongoing research (e.g. Van Dongen, 2006; Lele and Dennis, 2009; Gelman et al., 2017). Our solution here was regularization of many parameters, for illustrative effect, but we encourage analysts thoroughly investigate the implications of, and justification for, regularization in real-world cases. Thus, despite our success here, it is unlikely “flipping the switch” results in fast run times for existing models because regularization is an instrumental and necessary part of Bayesian inference in stock assessments.

The necessary steps for regularization will vary by model, but we have demonstrated some powerful tools to help guide this process. First, we note that Bayesian modelling involves a process where, among other steps, complexity should be slowly increased (Gelman et al., 2014; Gabry et al., 2019). Here, we do the opposite because we expect existing models to start overparameterized and

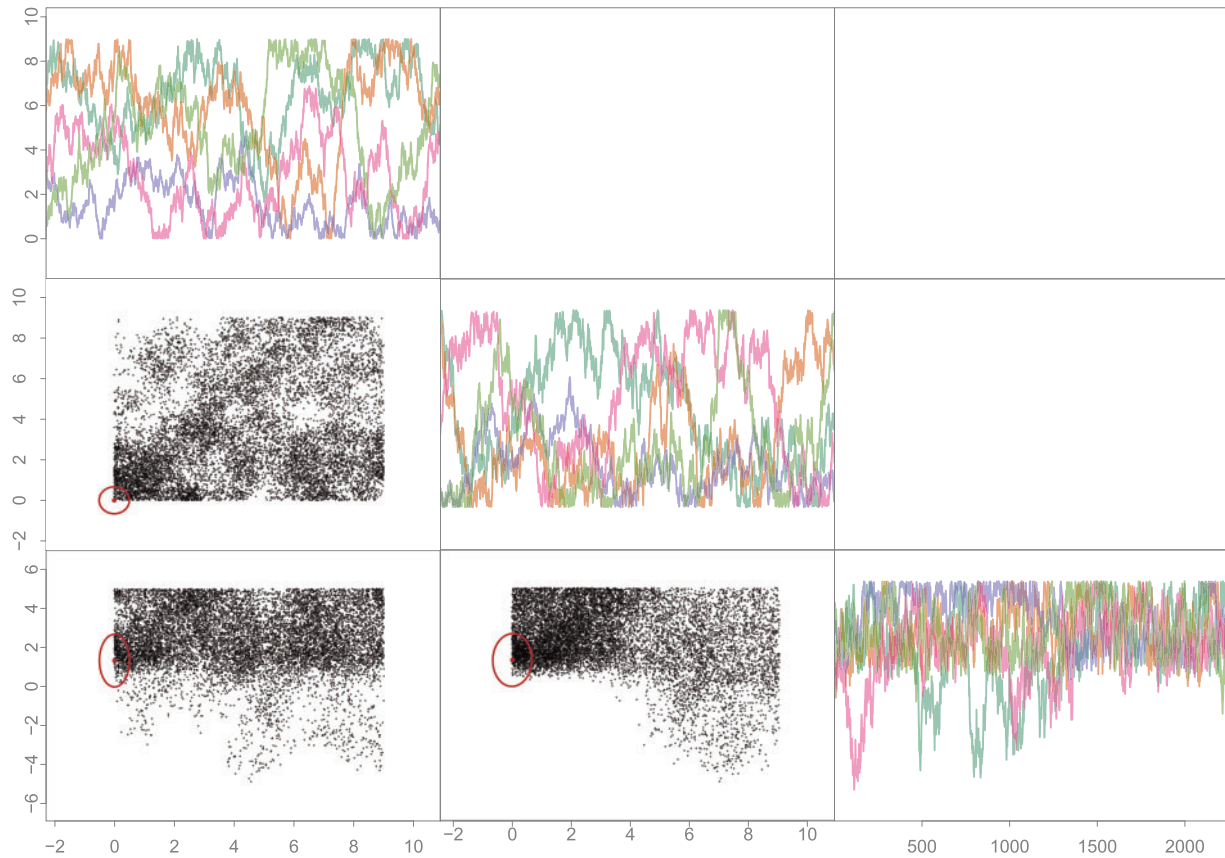


Figure 2. Results from pilot chains for the Canary model. Shown are three selectivity parameters, which mixed particularly poorly because of flat regions of the posterior. See Figure 1 caption for interpretation.

the goal is to constrain them. This complexity in assessment models is often justified by authors because many versions of the assessment are run during scientific review. In this context, parameters poorly informed for one configuration may not be for another, so it is “safer” to estimate them all, assuming no negative effects on frequentist inference. Further, overly regularized models may fail to represent legitimate uncertainty (alternate hypotheses about the population dynamics), and therefore lead to incomplete information for fisheries management. However, caution should be taken with adding too much complexity with frequentist inference as well (Subbey, 2018). Regularization, whether fixing parameters or adding explicit priors, is a key but difficult step in constructing a Bayesian assessment and we encourage analysts to thoroughly and thoughtfully explore how to best do this process. We also note that the goal is to arrive at a model, which is parameterized to be commensurate with the information in the data and priors while also being useful for management purposes. With that in mind, to adapt an existing model for Bayesian inference we recommend the subsequent steps in conjunction with the `adnuts` R package (Monnahan, 2018), which streamlines the workflow and provides additional features not available to command line execution:

- (i) Incorporate all available informative priors on model parameters.
- (ii) Run parallel pilot RWM chains (at least three times) started from parameter MLEs. A good starting place is

chains long enough to obtain 1000 samples after thinning every 100 and 20% warmup, but a lower thinning rate for slow models may be appropriate.

- (iii) Identify slow mixing parameters using diagnostic plots (e.g. Figures 2 and 3) and regularize or reparametrize as appropriate.
- (iv) Rerun pilot chains and compare frequentist estimates of key management quantities to previous runs (e.g. Table 2).
- (v) Repeat steps 2–4 until all parameters are mixing at a reasonable rate.
- (vi) Run parallel pilot chains with NUTS from previous posterior draws, producing 500 samples with no thinning. If divergences exist, identify the cause. Solutions include more regularization or reparameterizing and increasing the target acceptance rate.
- (vii) Run inference chains using NUTS with 2000 samples, 200 warmup iterations, and no thinning, using updated mass matrix estimated from the first NUTS analysis, which improves efficiency but returns the same posterior.
- (viii) Check for lack of non-convergence and lack of divergences, then use these samples for inference.

We recommend the RWM algorithm for exploratory analysis because it worked better than NUTS in the presence of grossly mismatched mass matrices, but the NUTS algorithm should be used for inference for two reasons. First, it was consistently faster for

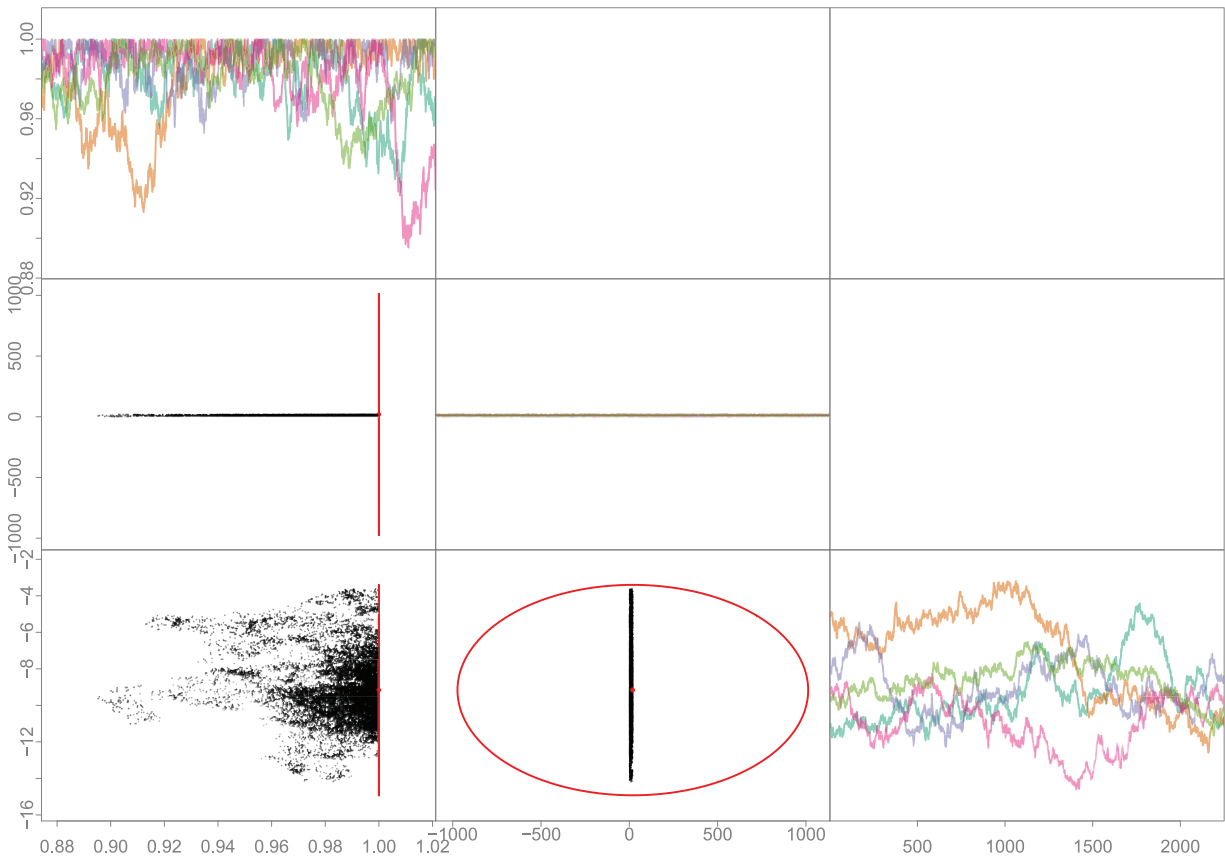


Figure 3. Results from pilot chains for the Snowcrab model. Shown are select parameters, which mixed particularly poorly because of an inverse Hessian that failed to accurately describe the underlying posterior, being either too small (first column) or too big (second column). See Figure 1 caption for interpretation.

the same model, particularly when an estimate of the mass matrix was available from a previous run (Table 3). Second, it was better able to explore extremely difficult geometries making it less susceptible to biased posteriors and additionally warned if bias may be occurring, whereas RWM does not do this (Figure 4; Betancourt, 2016; Monnahan *et al.*, 2017). We believe NUTS is a valuable tool for analysts performing Bayesian inference on stock assessments in ADMB and recommend it as the default algorithm for inference.

Our work here paves the way for future research on the use of Bayesian inference for data-rich stock assessments, and we highlight several avenues for fruitful extensions. First, the double-normal selectivity is particularly challenging for MCMC samplers, at least with the default priors found in most of the models. Development of flexible selectivity parameterizations that are more commensurate with Bayesian integration in addition to maximum likelihood should be investigated. One intriguing option is a non-parametric or semi-parametric approach (Thorson and Taylor, 2014; Xu *et al.*, 2018). Additionally, random effects are a flexible and powerful tool for modelling various biological and fisheries processes, but their hypervariances are currently inestimable in ADMB and they are typically fixed at a constant value (although see Thorson *et al.*, 2015a). We thus encourage investigations into the estimability of one or more hypervariances, and note that NUTS is particularly promising because it is efficient for complex mixed effects models (Betancourt and Girolami, 2015; Monnahan *et al.*, 2017). One technical difficulty

in ADMB is our proposed workflow uses the inverse Hessian as the mass matrix (Supplementary Appendix A), but this matrix may not be defined in a hierarchical model. One solution would be dense mass matrix adaptation as part of the NUTS warmup period, as done in Stan but not currently available for ADMB (Supplementary Appendix A; Stan Development Team, 2017). Clearly, this would be an important addition to ADMB, whether hypervariances are estimated or not, as it would improve overall speed and simplify the Bayesian workflow. Finally, we recommend further investigation into the causes of differences (or lack thereof) in inference between the two statistical paradigms (Figure 6).

Assessment models are often adapted and improved within a management framework of scientific review panels, sometimes requiring updated results overnight. Given its speed advantage, it is not surprising that maximum likelihood estimation is the predominant method for inference, whereas long Bayesian run times are an obstacle. Here, we showed that orders of magnitude improvements in Bayesian run time could be achieved with regularization, faster algorithms, and parallel chains. Still, it takes concerted effort to regularize a model, and run times are still much slower than frequentist estimation. Despite this, we argue that Bayesian inference still provides value for at least three reasons. First, it provides a formal way to incorporate prior information and evaluate the consequences of alternative management actions for use in a decision analysis (Punt and Hilborn, 1997). Second, it helps diagnose structural issues of a

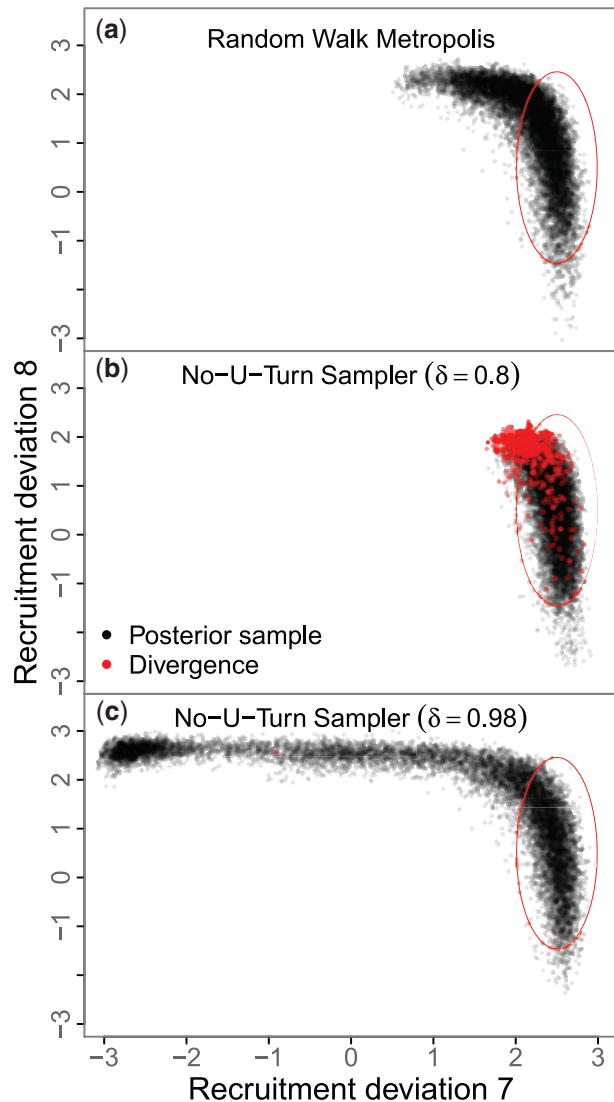


Figure 4. Posterior samples from two adjacent recruitment deviations in the regularized Halibut model with the priors for these two parameters removed (i.e. partially regularized). (a) Is from the RWM algorithm whereas (b) and (c) are the NUTS for two different levels of the target acceptance rate (δ , and called `adapt_delta` in `adnuts`). Black points are posterior samples, whereas red points are divergent transitions (only outputted by NUTS, see online color version). In (a) and (b) the algorithms cannot generate samples from a subset of the posterior, which leads to biased estimates of these parameters, but in (b) NUTS with the default of $\delta = 0.8$ warns about this potentially bias with divergences. By increasing δ to 0.98 the algorithm eliminates divergences and samples from a wider region of the posterior. The ellipse shows the 95% credible interval for the prior assigned to these parameters during regularization, but which is not applied here but eliminated these sampling difficulties in the final model with minimal effect on management quantities.

model, which are undetectable just looking at MLEs and covariances. We uncovered violations of the frequentist assumptions in real stock assessments and highlight that despite these violations, ADMB still produced an invertible Hessian matrix, suggesting this is insufficient evidence a model meets its assumptions. Because of the unique perspective Bayesian

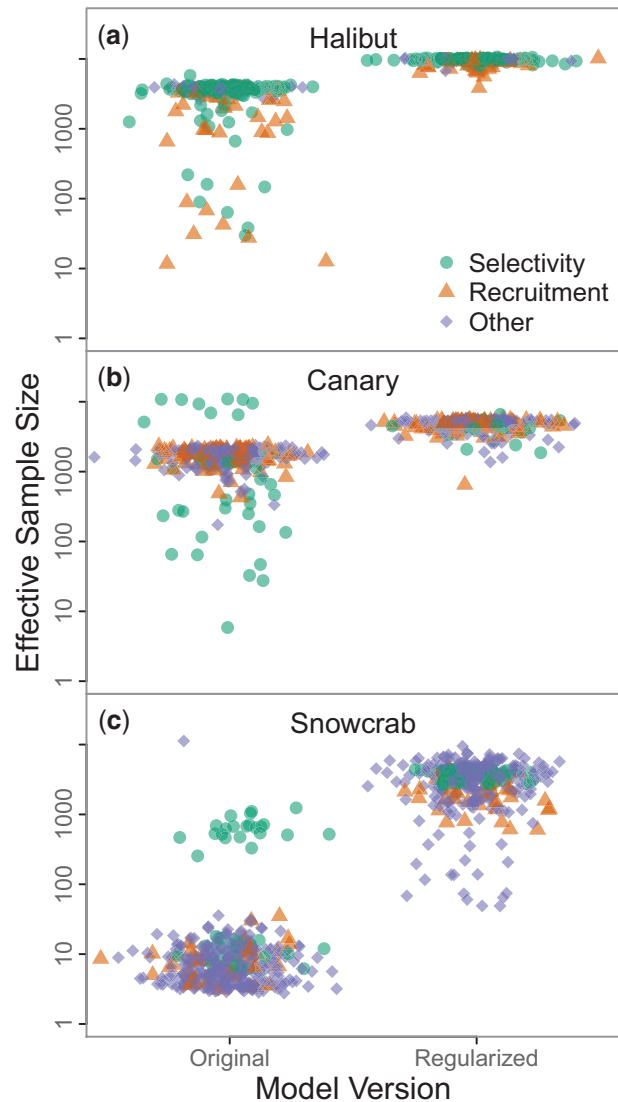


Figure 5. Improvement through regularization for the RWM pilot chains. Effective sample sizes estimated before and after regularization using five chains with 3000 samples after thinning every 1000th sample and discarding the first 25% as a warmup period (thus 7500 nominal samples). The smallest effective sample size defines the mixing rate and thus efficiency of the chains. The parameter type is differentiated by point colour and shape. x-axis values are jittered for visual clarity.

integration gives into a model, we argue it is a useful tool for developing more statistically robust models in the frequentist paradigm. Finally, the choice of paradigm may or may not lead to different management advice (Figure 6; Stewart *et al.*, 2013), although this is not known *a priori*. We do not argue that one paradigm is better, but rather highlight the value in comparing inference between the two methods on the same stock assessment. Fortunately, combining powerful Bayesian integration via NUTS and ever-increasing computational power, reasonable run times are possible even for the largest, slowest fisheries assessment models. This opens the door to future research to improve stock assessment models and, where desirable, applied inference in management scenarios.

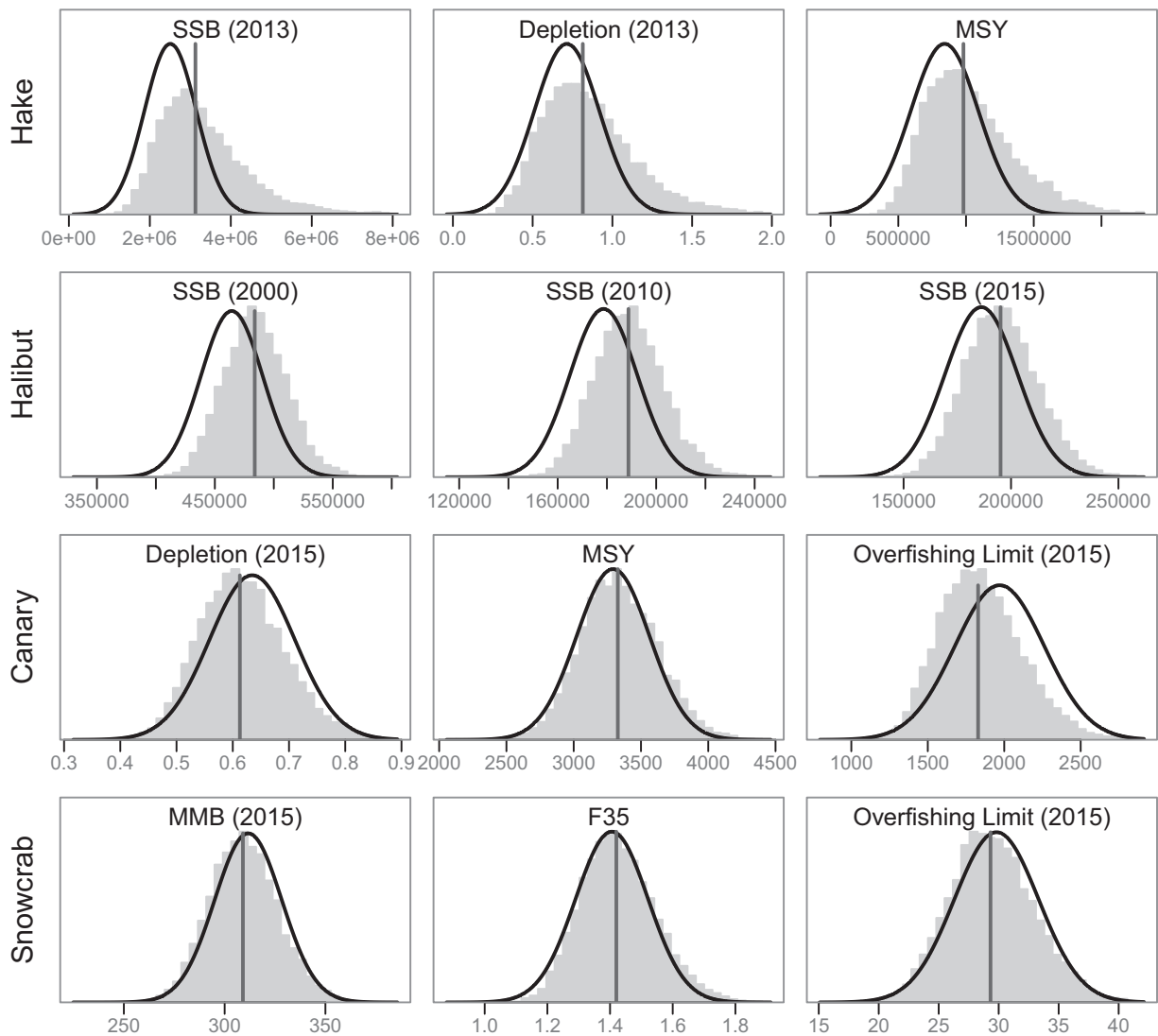


Figure 6. Comparison of estimates of relevant management quantities for the regularized models (Table 2) between frequentist and Bayesian paradigms. Posterior distributions are shown as grey histograms with vertical line and point showing the median. Asymptotic estimates from the delta method, assumed to be normally distributed, are shown as black curves, and the posterior median is shown as a vertical line. SSB is spawning-stock biomass, MMB is mature male biomass, depletion is SSB relative to unfished, MSY is maximum sustainable yield, and F35 is the overfishing level fishing mortality.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

We thank Kelli Johnson and two anonymous reviewers for helpful feedback on a previous draft. We also thank Richard Methot for providing Stock Synthesis ADMB code and general discussion, as well as Ian Taylor and Allan Hicks for general discussion of Stock Synthesis and the Hake model.

Funding

This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA15OAR4320063, Contribution No. 2018-0171. This work was partially funded in part by a grant

from Washington Sea Grant, University of Washington, pursuant to National Oceanic and Atmospheric Administration Award No. NA14OAR4170078. TAB was also funded in part by the Richard C. and Lois M. Worthington Endowed Professorship in Fisheries Management. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

References

- Betancourt, M. 2016. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. arXiv preprint arXiv:1604.00695.
- Betancourt, M. 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434.
- Betancourt, M., and Girolami, M. 2015. Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, 79: 30.

- Cotter, A. J. R., Burt, L., Paxton, C. G. M., Fernandez, C., Buckland, S. T., and Pan, J. X. 2004. Are stock assessment methods too complicated? *Fish and Fisheries*, 5: 235–254.
- de Valpine, P. 2009. Shared challenges and common ground for Bayesian and classical analysis of hierarchical statistical models. *Ecological Applications*, 19: 584–588.
- Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., and Nielsen, A. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27: 233–249.
- Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182: 389–402.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2014. *Bayesian Data Analysis*. vol. 2. Chapman, Boca Raton, FL.
- Gelman, A., Simpson, D., and Betancourt, M. 2017. The prior can often only be understood in the context of the likelihood. *Entropy*, 19: 555.
- Grandin, C. J., Hicks, A. C., Berger, A. M., Edwards, N., Taylor, I. G., and Cox, S. 2016. Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2016. Prepared by the Joint Technical Committee of the U.S. and Canada Pacific Hake/Whiting Agreement, National Marine Fisheries Service and Fisheries and Oceans Canada. 165 pp.
- Hilborn, R., and Walters, C. J. 1992. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*. Norwell, Massachusetts, USA.
- Hoffman, M. D., and Gelman, A. 2014. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15: 1593–1623.
- Kuriyama, P. T., Ono, K., Hurtado-Ferro, F., Hicks, A. C., Taylor, I. G., Licandeo, R. R., Johnson, K. F., *et al.* 2016. An empirical weight-at-age approach reduces estimation bias compared to modeling parametric growth in integrated, statistical stock assessment models when growth is time varying. *Fisheries Research*, 180: 119–127.
- Lele, S. R., and Dennis, B. 2009. Bayesian methods for hierarchical models: are ecologists making a Faustian bargain? *Ecological Applications*, 19: 581–584.
- Link, W. A., and Eaton, M. J. 2012. On thinning of chains in MCMC. *Methods in Ecology and Evolution*, 3: 112–115.
- Magnusson, A., Punt, A. E., and Hilborn, R. 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. *Fish and Fisheries*, 14: 325–342.
- Maunder, M. 2003. Paradigm shifts in fisheries stock assessment: from integrated analysis to Bayesian analysis and back again. *Natural Resource Modeling*, 16: 465–475.
- Maunder, M. N., and Punt, A. E. 2013. A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142: 61–74.
- Method, R. D. 2015. User Manual for Stock Synthesis. Version 3.24s. http://www.st.nmfs.noaa.gov/Assets/science_program/SS_User_Manual_3.24s.pdf (last accessed 16 June 2015).
- Method, R. D., and Wetzel, C. R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, 142: 86–99.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21: 1087–1092.
- Monnahan, C. C. 2018. adnuts: No-U-Turn Sampling for ADMB and TMB Models. R Package Version 1.0.1. <https://cran.r-project.org/package=adnuts>
- Monnahan, C. C., and Kristensen, K. 2018. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: introducing the adnuts and tmbstan R packages. *PLoS One*, 13: e0197954.
- Monnahan, C. C., Ono, K., Anderson, S. C., Rudd, M. B., Hicks, A. C., Hurtado-Ferro, F., Johnson, K. F., *et al.* 2016. The effect of length bin width on growth estimation in integrated age-structured stock assessments. *Fisheries Research*, 180: 103–112.
- Monnahan, C. C., Thorson, J. T., and Branch, T. A. 2017. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8: 339–348.
- Neal, R. M. 2011. MCMC using Hamiltonian dynamics. *In Handbook of Markov Chain Monte Carlo*, Ed. by S. Brooks, A. Gelman, G. L. Jones, X.-L. Meng. CRC Press in Boca Raton, FL, USA. p. 113.
- Punt, A. E., and Hilborn, R. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. *Reviews in Fish Biology and Fisheries*, 7: 35–63.
- Quinn, T. J., and Deriso, R. B. 1999. *Quantitative Fish Dynamics*. Oxford University Press.
- R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Sampson, D. B., Scott, R. D., and Quinn, T. 2011. A spatial model for fishery age-selection at the population level. *Canadian Journal of Fisheries and Aquatic Sciences*, 68: 1077–1086.
- Schnute, J. 1981. A versatile growth model with statistically stable parameters. *Canadian Journal of Fisheries and Aquatic Sciences*, 38: 1128–1140.
- Stan Development Team. 2017. Stan Modeling Language Users Guide and Reference Manual. Version 2.17.0. <https://mc-stan.org/users/documentation/>.
- Stan Development Team. 2018. rstan: R Interface to Stan. R Package Version 2.18.2. <http://mc-stan.org>.
- Stewart, I. J., Hicks, A. C., Taylor, I. G., Thorson, J. T., Wetzel, C., and Kupschus, S. 2013. A comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. *Fisheries Research*, 142: 37–46.
- Stewart, I. J., and Martell, S. J. D. 2015. Reconciling stock assessment paradigms to better inform fisheries management. *ICES Journal of Marine Science: Journal du Conseil*, 72: 2187–2196.
- Stewart, I. J., and Monnahan, C. C. 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. *Fisheries Research*, 192: 126–134.
- Stewart, I. J., Monnahan, C. C., and Martell, S. 2016. Assessment of the Pacific Halibut Stock at the End of 2015. IPHC Report of Assessment and Research Activities, pp. 188–209. <https://iphc.int/uploads/pdf/rara/iphc-2015-rara25.pdf> (last accessed 26 March 2019).
- Subbey, S. 2018. Parameter estimation in stock assessment modelling: caveats with gradient-based algorithms. *ICES Journal of Marine Science*, 75: 1553–1559.
- Szuwalski, C. S. 2016. Biases in biomass estimates: the effect of bin width in size-structured stock assessment methods. *Fisheries Research*, 180: 169–176.
- Szuwalski, C. S., and Turnock, J. 2016. A stock assessment for eastern Bering Sea snow crab. https://www.npfmc.org/wp-content/PDFdocuments/resources/SAFE/CrabSAFE/2018/1-EBSSnow_SAFE_2018.pdf (last accessed 26 March 2019).
- Thorson, J. T., and Cope, J. M. 2017. Uniform, uninformed or misinformed? The lingering challenge of minimally informative priors in data-limited Bayesian stock assessments. *Fisheries Research*, 194: 164–172.
- Thorson, J. T., Hicks, A. C., and Method, R. D. 2015a. Random effect estimation of time-varying factors in Stock Synthesis. *ICES Journal of Marine Science*, 72: 178–185.
- Thorson, J. T., and Minto, C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. *ICES Journal of Marine Science*, 72: 1245–1256.

- Thorson, J. T., Monnahan, C. C., and Cope, J. M. 2015b. The potential impact of time-variation in vital rates on fisheries management targets for marine fishes. *Fisheries Research*, 169: 8–17.
- Thorson, J. T., and Taylor, I. G. 2014. A comparison of parametric, semi-parametric, and non-parametric approaches to selectivity in age-structured assessment models. *Fisheries Research*, 158: 74–83.
- Thorson, J. T., and Wetzel, C. 2015. The Status of Canary Rockfish (*Sebastes pinniger*) in the California Current in 2015. http://www.cio.noaa.gov/services_programs/prplans/pdfs/ID308_FinalProduct_CanaryRockfish_2016.pdf (last accessed 26 March 2019).
- Van Dongen, S. 2006. Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242: 90–100.
- Xu, H., Thorson, J. T., Methot, R. D., and Taylor, I. G. 2019. A new semi-parametric method for autocorrelated age- and time-varying selectivity in age-structured assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 76: 268–285.

Handling editor: Shijie Zhou