Draft Report of the Research Track Working Group for Applying State-space models

January 28, 2024

Table of Contents

Executive Summary

A research track “applying state space assessment methods for fisheries management” was undertaken and initially planned for peer review in November 2023, but was postponed until February 12-15, 2024. The Working Group (WG) was formed in August 2021 and met regularly to address its terms of reference (TORs). This report represents the consensus of the WG and includes contributions from WG members and external collaborators.

TOR 1: Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.

The WG addressed this ToR through review of: 1) the scientific literature on state-space modeling, 2) the scientific literature on state-space stock assessment modeling, 3) the scientific literature on assessment model diagnostics, and 4) of relevant results from working papers prepared by members of the WG. These reviews formed the basis of recommended practices for selecting among alternative configurations of state-space stock assessment models. Recommendations by the WG are:

1. Treat recruitment as random effects so that variance and correlation parameters can be estimated
2. Consider as many sources of process error as might be plausible and practical, but be aware of unintended implications for management reference points and catch advice.
3. When non-negligible mis-reporting of catch is plausible, estimation of catch process errors should be considered.
4. When reliable external estimates of observation error are available treat them as known in the assessment model.
5. Perform posterior check of all random effects.
6. When using MASE with time-series cross-validation, we recommend using the denominator as described by Hyndman and Koehler (2006). A generalization of MASE using (randomized) quantile prediction errors is needed.
7. Use a broad suite of metrics and diagnostic tools to evaluate relative performance of alternative models. Statistical reliability and AIC as a model selection tool are better when there is contrast in fishing pressure, stock size and process errors over time and more precise index and age composition observations are available.
8. For Bayesian fitting methods, do not use DIC as a criterion for determining a preferred model.

TOR 2: Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.

The WG addressed this ToR through: 1) review of the best available science on stock-recruit modeling, and 2) analyses of simulation studies carried out by the WG. These reviews formed

the basis of recommendations on scenarios where inferences about stock-recruit relationships are most reliable and diagnostics to use to increase confidence in reliability of results.
Recommendations by the WG are:

1. Consider the level of information in the stock assessment data for the stock-recruit relationship. Positive responses to these questions increase the likelihood for reliable inferences
    a. ***Is the time series sufficiently long?***
    b. ***Is there evidence of good contrast in spawning stock biomass over time?***
    c. ***Are index and age composition observations relatively precise?***
    d. ***Is variation in recruitment residuals (sigma-R) relatively low?***
2. Estimate the stock-recruit relationship simultaneously and internal to the state-space stock assessment model.
3. Self-tests as described in TOR 1 would be prudent to confirm reliability of stock-recruit parameter estimates and biological reference points derived from them.
4. Consider alternative autocorrelation models for recruitment residuals. This will be important primarily in defining how recruitment is predicted in short-term projections.

TOR 3: Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.

The WG addressed this ToR through: 1) review of the best available science on modeling environmental effects in stock assessment models, and 2) analyses of simulation studies carried out by the WG.
Recommendations by the WG are:

1. Limit investigations to covariates that biology suggests close links of the covariate to the particular demographic parameter.
2. Evaluate effects of covariates against models that have temporal variation in the parameter which the covariate is hypothesized to affect.
3. Check whether observation error in environmental covariates observation is low relative to other data sources as this improves reliability of inference and estimability.
4. Fix parameters describing environmental process variability where information is known.
5. Avoid the ‘masking’ functional form when relating stock-recruitment relationships to an environmental covariate (until further work can diagnose issues).
6. Ensure good contrast in the environmental covariate(s).
7. Conduct retrospective comparisons of models with and without covariate effects to confirm inferences are consistent as the number of years with observations changes.
8. Conduct self-tests as described in TOR 1 to confirm reliability of the estimation of effect size the covariate has on assessment model parameter estimates and reliability of biological reference points.

TOR 4: Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history

type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.

The WG was unable to complete simulation studies targeted to this ToR. The WG reviewed relevant previous research and described a simulation study design that could be used addresse this ToR.

TOR 5: Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

External collaborators and working group members undertook modeling of four stocks using the WHAM package for state-space age structured stock assessment models. The four stocks and assessment leads are

- Georges Bank winter flounder: Alex Hansell
- Atlantic mackerel: Kierstin Curti and Alex Hansell
- Acadian redfish, Brian Linton
- Gulf of Maine haddock: Charles Perretti

The leads on the respective working papers completed bridge model runs from the current assessment modeling platform to WHAM with limited application of random effects as a proposed model to move to the management track where further exploration of process errors can be conducted using guidance and recommendations from this research track with further peer-review.

Working group process

The WG (Appendix 1) initially met August 23, 2021 and scheduled 33+ semi-weekly meetings from February 2022 until December 2023 and weekly through the end of January 2024. Meetings were canceled when there was no progress to report or the chair was unavailable. Meetings were all held virtually to facilitate attendance and to be open to the public (Appendix 2), with information on how to attend provided on the dedicated website for the working group (https://www.fisheries.noaa.gov/event/applying-state-space-models). For each meeting, the chair prepared an agenda and requested items from working group members and other external participants. Over the course of the period that the WG met, we solicited presentations from external experts on state-space assessment methods and investigators conducting research on state-space assessment methods. The WG decided to meet the terms of reference by both reviewing the relevant literature and conducting several large scale simulation studies. The WG made extensive use of Google documents, a Github repository (https://github.com/timjmiller/SSRTWG), and high-performance computing resources at the University of Massachusetts and Woods Hole Oceanographic Institution and the MS Azure cloud computing platform. The usage of these computing resources was made possible by members of the WG and colleagues at NOAA Fisheries HQ. Without these external resources the WG would not have been able to conduct the work needed for this research track peer-review. Our only option to house files containing the results from all of the simulation studies (>200GB) and allow access to external working group members and participants was Google Drive. The completion of the report also benefited greatly from contributions from several collaborators external to the WG.

Introduction

What are state-space models?

State-space models combine time series models of latent processes and periodic observations (or measurements) of this process that include error. Assuming a state-space model with latent processes and measurements both being linear and Gaussian, Kalman (1960) introduced a set of equations for calculating optimal estimates of latent processes given known covariances of the latent process and observations: the Kalman filter. A wide array of Gaussian time-series models can be portrayed in state-space form to take advantage of the Kalman filter (Durbin and Koopman 2001). Although classical state-space models assume multivariate Gaussian latent processes and Gaussian observations at each time point, generalizations to non-Gaussian latent processes and observation and non-linear conditional expectations of the latent process or the observations of the latent process exist (Durbin and Koopman 2001). State-space models are a special type of a larger class of hidden process models (Markovian) where the distribution of the state at time *t*, given the state at time *t*-1, is known (Newman et al 2006). Viewing the latent states as random effects, state-space models are also a special type of mixed effects model.

The parameters of state-space models can be estimated by maximum likelihood or using fully Bayesian methods. In the former, the marginal likelihood integrated over any random effects is maximized with respect to the fixed effects parameters. Then the random effects are parametric empirical Bayes estimates given the MLEs for the fixed effects. In the latter, prior distributions would be assumed for all of the fixed effects parameters as well. Both approaches are also used in more widely known linear or generalized linear mixed effects models (e.g., Fong et al. 2010, Brooks et al. 2017) .

An important nuance of using state-space models in practice is whether the variance parameters associated with latent stochastic processes and those associated with the observation measurement error are estimated. Estimating variance parameters for both the latent process and the observations is known to be challenging (Dennis et al. 2006, Knape 2008, Fay and Punt 2013, Auger-Méthé et al. 2016).

The more recent combination of automatic differentiation software and numerical integration has allowed more rapid estimation of state-space models. The random effects module of ADMB (Fournier et al. 2012) and, more recently, the Template Model Builder package (TMB, Kristensen et al. 2016) in R has made rapid development and fitting of complex non-linear, non-Gaussian state-space models more approachable. These packages allow maximum marginal likelihood estimation of all fixed effects including latent process variance parameters by employing the Laplace approximation of the marginal log-likelihood and its gradient to be evaluated. However, fully Bayesian estimation of all parameters is possible by joining TMB models with STAN (RSTAN, Stan Development Team 2024; TMBSTAN, Monnahan and Kristensen 2018).

State-space models in fisheries stock assessment

Fisheries stock assessment modeling was one of the earliest ecological fields to explore state-space approaches (Auger-Méthé et al 2021). Walters and Hilborn (1976) used the Kalman filter to update Ricker stock-recruit and Schaefer surplus production model parameters as new annual observations arise. Mendelssohn (1988) and Gudmundsson (1994) demonstrated the use of the Kalman filter for modeling transitions in population abundance at age and annual index observations. Sullivan (1992) showed how to use the same approach for length structured models. Since then, complexity in assessment models has grown with the advances in numerical methods and computational capacity, including allowing many assessment parameters to vary through time. For example, the Age Structured Assessment Program, ASAP (Legault and Restreppo 1999) has as an option to allow catchability to vary through time as a random walk process. Stock Synthesis (Methot and Wentzel 2013) allows many parameters to vary temporally. However, these models have consistently treated the time-varying latent parameters as fixed effects rather than random effects with variance parameters for the time-varying processes assumed known and estimation done via penalized maximum likelihood.

In the last decade, development of state-space assessment models using ADMB and TMB has allowed estimation of both random and fixed effects, including variance parameters for latent processes and observations. The computational advances have made it practical to use these models to manage fish stocks, determining stock and fishing status, and making catch advice. The State-space stock Assessment Model (SAM; Nielsen and Berg 2014) was the first general state-space age structured model to be used to manage fish stocks. It is now widely used for stocks managed by the International Council for the Exploration of the Sea (ICES). Cadigan (2016) developed a state-space age structured model that is used for management of Northern cod in Canada. Miller et al. (2016) developed a state-space model that allows latent time series processes for the population as well as environmental covariates that may affect recruitment of the population. The model was generalized and built into a software package in R known as the Woods Hole Assessment Model (WHAM) by Stock and Miller (2021).

Although the statistical aspects of state-space models and their application have been studied extensively, the work is primarily on Gaussian state-space models. The application to stock assessments is in many ways the most complex implementation of the state-space approaches to parameter estimation. The equations for transitions of the latent processes are typically non-linear as are the equations relating the processes to observations. Furthermore, there can be several observation sources in assessment models with different probability distributions.

Current consensus is that use of state-space methods is recommended as best practice and should be used in next-generation stock assessment packages (Hoyle et al. 2020, Hoyle et al. 2022, Punt 2023). State-space methods provide improved statistical rigor in estimating time-varying processes and better representation of uncertainty in assessment model output.

The Woods Hole Assessment Model

Because WHAM is the primary state-space assessment software already used for management in the Northeast US (NEUS) and it is used for all of the research and applications conducted as part of this research track, we provide here a description of the software package as it currently exists. Further details, examples, and demonstrations can be found in the vignettes for the package: https://timjmiller.github.io/wham/articles/index.html.

The WHAM package uses TMB to create and fit models using the "nlminb" optimizer in R. The package is open source and freely available as a github repository (https://github.com/timjmiller/wham). WHAM is under active development with new features being added and tested. There have been 8 releases to date and two major extensions to model 1) length composition and growth (Correa et al. 2023) and 2) multiple stocks and regions with movement are intended to be combined in the standard WHAM software in the near future.

The current version of WHAM includes options to include random effects on recruitment, the transitions in abundance at age between time steps ("survival"), natural mortality, selectivity parameters, index catchability, and environmental covariates. As in Miller et al. (2016), and Miller et al. (2018), environmental covariates are treated in a state-space framework as well because the annual covariates are typically derived from models or calculated from samples of measurements and therefore are an estimate (with error) of the true covariate which is considered a latent time series process. When covariates are included, they may or may not be assumed to affect recruitment, natural mortality, and/or index catchability.

Recruitment and survival: When annual recruitments are treated as random effects, there is a mean recruitment and annual recruitments can either be assumed independent or to be first order autoregressive (AR1). Survival random effects are log-normal and can be assumed independent, AR1 across age (and independent across years), AR1 across years (and independent across age), or AR1 across age and year (2DAR1). By default, recruitment is treated as a random effect when survival is and any correlation across age applies to both recruitment and survival random effects. However, options exist in the developmental unreleased version to decouple correlation of recruitment and survival.

Selectivity: Selectivity parameters are estimated on a logit scale and when random effects are assumed, they are Gaussian on the logit scale. Options are to apply: a AR1 yearly random effect that is the same for all selectivity parameters in a given year; AR1 random effects across parameters that are the same for each years; and random effects for each selectivity parameter each year that may be independent or have a 2DAR1 structure in year and across parameters within year. Mean or constant selectivity can be increasing or decreasing logistic functions of age, double logistic, or as age-specific parameters. When age-specific mean parameters are assumed, random effects are only allowed for parameters that are not fixed at 1 or 0.

Natural Mortality: Like survival random effects, when random effects are assumed for natural mortality, they are Gaussian on log scale. Similar AR1 and 2DAR1 correlation structures as selectivity are options for natural mortality random effects.

Catchability: The catchability of each index can be assumed to vary over time. Catchability is estimated on a logit scale with bounds 0 and 1000 by default and annual random effects are Gaussian on this scale. The annual catchability random effects can be assumed independent or AR1 in nature.

Environmental covariates: Any number of environmental covariate time series can be included as latent random effects. The latent covariates are assumed to be independent and each time series can be assumed to have a simple random walk or a stationary AR1 autocorrelation as in the above AR1 assumptions above. Effects of covariates are linear on some transformed scale that depends on what parameter the covariate is affecting. However, there are options to assume orthogonal polynomial effects of each covariate and the time lag of when the covariate affects the population can be specified.

Environmental effects on recruitment: Effects can be assumed whether or not there is a stock recruit relationship (SRR) assumed. Without the SRR, the effect of covariates is linear on log recruitment. When a Beverton-Holt stock recruit relationship is assumed, there are multiple options for mechanistic hypotheses of effects on the relationship (Iles and Beverton 1999). There are also two options for mechanistic hypotheses when a Ricker SRR is assumed (See Stock and Miller 2021 for further details).

Environmental effects on natural mortality: When effects of covariates on natural mortality are assumed, they are linear on log natural mortality and can be assumed for specific ages or in combination with allometric functions of weight at age.

Environmental effects on catchability: As index catchability is estimated on a logit scale any covariate effects are linear on this logit scale.

Data: Much of the structure of the input data for WHAM is similar to that of ASAP. There is functionality in WHAM to read ASAP input files and use default settings to create a WHAM input for a traditional statistical catch at age model similar, if not identical, to ASAP. In the NEUS, this is useful for bridging from one modeling framework to another in the research track process. The only new data type used in WHAM is observations of environmental covariates which are included and along with configuration of assumptions using the R functions of the WHAM package.

Simulation: WHAM can also be used to simulate any random effects and/or observations given defined parameters. The parameters can be defined from a fitted model or by the user without fitting a model as done in simulation studies completed by this working group.

WHAM use in management

WHAM has been approved to provide management advice for at least 10 fish stocks in the NEUS (Table 1.1). Most applications treat recruitment and survival as random effects, but fleet selectivity is also assumed to vary over time for Georges Bank and Eastern Georges Bank haddock, American plaice, and black sea bass.

Table 1.1 List of current assessment models using the WHAM package and types of process errors included as random effects

Year	Stock	
2022	Atlantic butterfish	Recruitment, Survival
2022	Georges Bank haddock	Recruitment, Survival, Fleet selectivity
2022	Eastern Georges Bank haddock	Fleet selectivity
2022	Atlantic bluefish	Recruitment, Survival
2022	American plaice	Recruitment, Survival, Fleet selectivity
2023	Eastern Gulf of Maine Atlantic cod	Recruitment, Survival
2023	Western Gulf of Maine Atlantic cod	Recruitment, Survival
2023	Georges Bank Atlantic cod	Recruitment, Survival
2023	Southern New England Atlantic cod	Recruitment
2023	Black sea bass	Recruitment, Survival, fleet selectivity, index selectivity, environmental covariate

Ongoing research track working groups for Golden tilefish (2024 peer review), three yellowtail flounder stocks (2024 peer review), and Atlantic herring (2025 peer review) are all developing assessment models using WHAM.

Motivation for this research track on applying state-space models

The Northeast U.S. Continental Shelf Large Marine Ecosystem (NES LME) encompasses Cape Hatteras, North Carolina in the south to the Gulf of Maine in the north and includes the management jurisdictions of both the Mid-Atlantic Fishery Management Council (NY-NC/VA border) and the New England Fishery Management Council (ME-CT) (Figure 1). This region is experiencing profound changes in physical and oceanographic properties as a result of both natural climate variability and human-induced climate change. The region is experiencing some of the fastest increases in average temperature observed globally with ocean temperatures increasing by 1.3°C since 1854 coupled with ocean acidification and increased rates of sea level rise (MAFMC 2019, Bates and Peters 2007). Climate projection models[1] predict continued increases in temperature, decreases in salinity, increases in precipitation, decreases in pH and continued sea level rise (Saba et al 2016, Yin et al 2013).

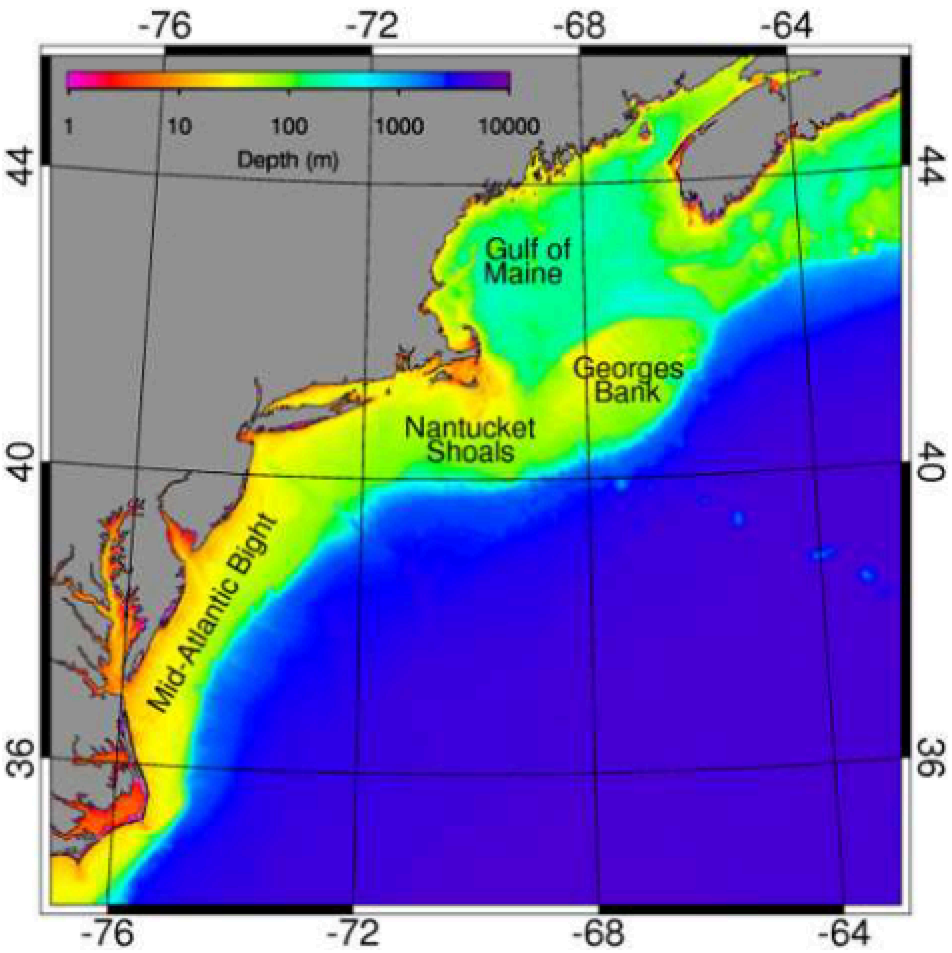

Figure 1. Bathymetric map of the Northeast Continental Shelf Large Marine Ecosystem**.**

[1] http://www.esrl.noaa.gov/psd/ipcc/ocn/

There are multiple potential biological responses to the pressures of climate change and variability. These pressures are likely to include changes in thermal habitat and hydrography, changes in primary productivity, changes in species composition, distribution, and interactions, and alterations to habitat scope and function. All of these, and others, are anticipated to have significant and different effects on the region's fisheries. Numerous studies have demonstrated long-term changes in the distribution and productivity of fish and shellfish resources on the Northeast U.S. Shelf (Hare et al. 2016). Changes in distribution have been documented in a large number of populations. Fewer studies have examined changes in stock productivity. The role of climate-forced changes in predator-prey dynamics needs to be investigated, but large-scale changes in species compositions suggests large-scale changes in predator-prey dynamics.

Important fisheries population dynamics parameters such as growth rates, natural mortality and recruitment are all likely to be vulnerable to climate change. Stock assessment models are highly dependent on these parameters. Stock assessment models, such as state-space models, that explicitly incorporate environmental drivers have the potential to adapt much more quickly and more accurately represent changing stock dynamics. Consequently, changes in these parameters will likely result in changes to biological reference points that are used in stock assessments and will have significant implications for the catch advice provided to the Councils. Given this direct application and consequences for management, there is high interest in exploring the potential application and increased utilization of state-space models within the region.

Management Councils and their Scientific and Statistical Committees (SSC), who set catch limits that Councils cannot exceed, are seeking more accurate assessment results and forecasts that incorporate contemporary conditions, such as those influenced by climate change and changes in ecosystem structure. SSCs are explicitly considering how stock-relevant ecosystem factors (e.g., environmentally dependent growth) are directly included in the assessment model and in short term projections when making allowable catch recommendations. They are increasingly looking for models that account for these changes and estimate uncertainty caused by environmental effects. Incorporation of environmental effects into assessment models has the potential to improve model performance when the current and future environment that affects stock productivity is different from historic conditions that are used by traditional assessment models to estimate stock productivity. The results of state space assessment models can then improve estimates of sustainable catch levels as well as measures that are necessary to rebuild overfished stocks.

Given the still fairly novel use of state-space methods in stock assessment and the various options available to the practitioner when applying these models for management, there is a clear need for guidance on whether certain options may be reliable or relevant and how to evaluate which of a set of alternative model configurations should be preferred.

Terms of Reference (TORs)

1. Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.
2. Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.
3. Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.
4. Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.
5. Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

Report Structure and Working Papers

The rest of this report is organized by sections for each of the TORs. In each section, the WG reviewed existing relevant literature and relevant results and conclusions from several working papers completed by members of the working group and external collaborators. At the end of each TOR section are recommendations and guidance from the WG for practitioners using state-space age structured assessment models and WHAM specifically. The 12 working papers are summarized in Appendix 3 and are available to the peer-review panel.

TOR 1: Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.

1.0 Introduction

There are several qualities that can be assessed to determine a preferred assessment model structure. These qualities include but are not limited to 1) better representation of realism of the biology and data generating mechanisms, 2) statistical reliability, 3) better prediction skill, 4) lack of evidence of model mis-specification. Many methods to assess these qualities are applicable whether or not the models are state-space in nature. Below we describe how these qualities might be assessed for state-space assessment models and, based on review of previous work and work conducted by the WG, provide recommendations on methods of evaluation and on future research to improve on these methods.

1.1 Realism

Assessment models with more biological realism have structures that represent the unique characteristics of the particular stock, such as spawning timing, growth, and maturation, and relationships to the environment, but those aspects are not specific to the state-space approach to assessment models. The primary way that state-space models may improve biological realism is by accounting for and estimating temporal variability in the demographic parameters that are otherwise treated as constant over time as well as the variance and autocorrelation of these processes.

1.1.1 Process errors to consider

Recruitment: There has been extensive work investigating temporal variation in many parts of age-structured stock assessment models, but recruitment is, by far, the most studied type of temporal variation in stock assessment models (Maunder and Deriso 2003, Maunder and Thorson 2019). Modeling temporal variation in SR parameters rather than deviations from a mean recruitment model has also been studied, although not within an integrated assessment model (Peterman et al. 2000, Britten et al. 2015). Within an integrated stock assessment model, treating the annual recruitments as random effects with marginal likelihood or Bayesian methods allows the variance and any autocorrelation to also be estimated rather than fixed or subject to likelihood profiling (Nielsen and Berg 2014, Cadigan 2016, Miller et al. 2016). The time series model used for recruitment will dictate how recruitment is predicted in short term projections and plays an important role in characterizing prediction uncertainty.

Survival Transitions: SAM and WHAM both allow random deviations from the expected numbers at age surviving between time steps (Nielsen and Berg 2014, Stock and Miller 2021). The expected survival is defined by the Baranov catch equations and these transitions are deterministic (no deviations) in traditional integrated catch at age models. WHAM allows these deviations to be independent or have AR1 structure across years and/or ages. Treating these survival transitions as random effects rather than the specific parameters allows the deviations to account for a wide variety of potential model mis-specifications including immigration and emigration, unspecified mortality sources and just any departures from the assumption of independent homogeneous mortality rates occurring continuously over the annual interval implicit in the Baranov equations. Including random effects on survival has been shown repeatedly to provide better model performance in the literature (Miller and Hyun 2018, Stock and Miller 2021) as well as in applications of WHAM in management.

Growth: Somatic growth has been shown to vary in space and time, and can be an important driver of fluctuations in productivity (Stawitz and Essington, 2018). Research on time-varying growth in stock assessment models has mostly been limited to penalized maximum likelihood models due to software limitations of the two most popular SSM frameworks, SAM and WHAM. For instance Kuriyama et al. (2016) used simulation testing to find that using an empirical weight at age approach is more accurate in data rich scenarios. Lee et al. (2018) explored linking an environmental covariate to parametric growth parameters and found improved performance in many cases, but that performance could degrade if the wrong covariate was used or ignored. Both of these studies used the Stock Synthesis framework (SS3; Methot and Wetzel 2013), which does not use marginal maximum likelihood. Using individual measurements of length, weight and age, Miller et al. (2018) showed the importance of including process errors in growth when evaluating evidence for temperature effects on growth and showed the effects on estimates of spawning stock biomass (SSB) and corresponding reference points of estimating this growth within a precursor to the WHAM model. WHAM was recently extended to have new approaches to estimating parametric and non-parametric growth variation, and to fit to marginal length data or conditional age-at-length data (Correa et al. 2023).

Selectivity: When there is variation in growth over time and fishing or survey gear are length selective, there will be variation in selectivity in models that treat selectivity as a function of age. Other mechanisms of changes in selectivity are temporal variability in 1) the composition of different gears operating in fishing fleets, 2) the legal size limits, 3) the availability of certain ages and 4) fleet targeting of certain ages (Martell and Stewart 2014, Sampson 2014). Time-varying selectivity has long been an option in traditional SCAA models by allowing constant selectivity parameters that differ between time periods (time blocks). Treatment of selectivity in age-structured stock assessments as a time-series process was considered as early as Gudmundsson (1994). In a simulation study examining performance of time-varying selectivity in Bayesian estimation models, Linton and Bence (2011) found better performance of models with time-varying double-logistic selectivity than models with age-specific selectivity parameters. Martell and Stewart (2014) suggest that when models parameterize selectivity as a function of age, time-varying selectivity parameters should be considered.

Catchability: Catchability is an important parameter in stock assessment because it is used to relate an index of abundance to stock size. Typically, stock assessments assume that catchability is constant; however, catchability can vary due to changes in sampling, the environment or ecology of the species (Maunder and Punt, 2004; Thorson et al. 2015). Incorrectly specifying catchability in a stock assessment can lead to biased trends in important management quantities. Most of the work looking at accounting for changes in catchability has been done outside of stock assessment models. For example, fisheries data is often standardized before being included in stock assessments (Hoyle et al. 2024). Additionally, fishery independent surveys are designed to have consistent sampling throughout time and space so catchability can be assumed constant. However, changing ecosystem dynamics can change the availability of fish to surveys and thus violate this assumption (Thorson 2019). To try to account for these changes, assessment models can use an environmental time series to help estimate time-varying catchability. Schirripa et al. (2017) found the Atlantic Multidecadal Oscillation (AMO) Index was correlated with Atlantic swordfish CPUE and using the AMO helped estimate catchability. Similarly, the AMO was used in the Atlantic bluefin tuna assessment to modulate time-varying catchability of United States and Canadian indices to account for a northward shift in the population (Hansell et al. 2020). It is difficult to fully understand the drivers of catchability so changes in catchability can also be included as a stochastic process. Liljestrand et al. (2023) used an autoregressive process to estimate age specific time-varying catchability for indices of abundance in a state-space stock assessment. Ultimately, many factors can influence catchability, which is why Wilberg et al. (2010) recommended that the default for assessment models should be to include time-varying catchability. Hart and Hansell (WP 2) simulation tested WHAM models with different approaches to time-varying catchability and found complementary results that models with environmental covariates characterize catchability well when environmental relationships are strong, well understood, and seasonally correct. However, if the environmental trend is not well understood or is seasonally misspecified, random effects are more effective at accounting for time-varying catchability.

Natural mortality: Natural mortality is most often treated as a constant value that is assumed known rather than estimated. However, there are numerous potential components of natural mortality such as predation, disease, starvation, and otherwise poor habitat, most of which can safely be assumed to vary seasonally and annually. Jiao et al. (2012) evaluated alternative assumptions about constant, and age and/or time-varying natural mortality for Atlantic weakfish using Bayesian fitting methods of an age structured model and found variation in both time and age provided the best fit. Deroba and Schueller (2013) found through a simulation study that accounting for temporal variation in natural mortality was more important for reducing bias in assessment output than accounting for differences at age. Johnson et al. (2015) evaluated the constant natural mortality assumptions that provided best estimation of assessment model output when true natural mortality was time varying. Trijoulet et al. (2020) demonstrated a state-space multi-species model that included observations of predation from stomach contents to inform natural mortality by year and age. Stock et al. (2021) demonstrated estimation of time- and age-varying natural mortality as random effects in WHAM for Southern New England and Mid-Atlantic (SNEMA) yellowtail flounder.

Missing catch: Under-reporting of landings is not an uncommon phenomenon. There are also situations where overreporting could occur. For example, if there are adjacent stocks of the same species and the quota is greater in one stock area than another, then overreporting in the high quota stock area and underreporting in the low quota area might be expected. This mis-reporting pattern could happen intentionally, or it could occur due to some seasonal/temporal overlap such that catch occurs in the overlap area but it all gets attributed to only one stock area. Another situation where overreporting could occur is the intentional mislabeling of catch of a low quota species (cod, e.g.) that is reported as a species that has a higher quota (haddock, e.g.), which occurred recently in the 'Codfather' case of Carlos Raphael in New Bedford, MA. When the fraction of total catch that is in error is large, there can be important impacts on the perception of the health of the stock and the catch advice. Cadigan (2016) was able to estimate both missing catch and natural mortality for Northern cod using a state-space model with constraints on the amount of missing catch. Perretti et al (2020) found little cost of estimating missing catch whether it is present or not using a state-space assessment model.

1.1.3 Observation error structure

Most stock assessment models that use observations of aggregate catch and indices assume a lognormal distribution for these observations (Methot and Wetzel 2003, Legault and Restreppo 1999). Current state-space assessment models also treat these observations as log-normal, or possibly multivariate log-normal observations at age (Nielsen and Berg 2014, Cadigan 2016, Stock and Miller 2021). Composition observations at length and/or age have been traditionally treated as multinomial random variables with some approach to determine an effective total sample size for each vector of observations (McAllister and Iannelli 1997, Francis 2011). However, more recently there has been a consensus that assessment models should move to self-weighting models with estimable dispersion parameters (Francis 2014, Cadigan 2016, Thorson et al. 2022).

Thorson et al. (2017) proposed an alternative parameterization of the Dirichlet-multinomial. The dispersion parameter scales with the input effective sample size. Simulation studies performed by Fisch et al. (2021), led the investigators to conclude that the logistic normal model for composition observations was more reliable, particularly with larger total sample sizes, than the Dirichlet-multinomial. However, simulation studies by Fisch et al (2023) led to a more nuanced conclusion that relative performance depended on the conditions of the operating model. Thorson et al. (2022) demonstrated a "multivariate" Tweedie treatment of frequencies at age (or length) which treats each of the frequencies as independent Tweedie random variables. The motivation was the equivalence of MLE estimation for multinomial observations treating the data as either multinomial or independent Poisson random variables. Factors that influence the reliability of the multivariate Tweedie remain an open question.

Liljestrand et al. (WP 12) simulated catch and index data under a range of age composition likelihood structures. Self- and cross- tests showed that correctly specifying the observation error structure reduced bias in terminal year spawning stock biomass (SSB) and exploitation rate. The logistic-normal and Tweedie distributions performed the best under the widest range of operating model scenarios but would over- or under- estimate the observation error variance, depending on how much process variability was specified in the process sub-model.

SAM treats observations of catch at age and abundance indices at age as multivariate log-normal (Nielsen and Berg 2014) so there is no need for a likelihood component for composition data. Albertsen et al. (2017) found this assumption to be preferred by AIC in applications of SAM to data on several stocks. However, when there are significant differences between the sampling methods used to estimate total catch or indices and the associated composition of ages or lengths, then it may be beneficial to use separate observation models for the aggregate and composition observations. WHAM has several options including the traditional multinomial, alternative Dirichlet-multinomial parameterizations, and alternative Dirichlet and logistic-normal configurations, and the multivariate-Tweedie.

When biomass is determined by growth within the assessment model, previous studies have found using conditional age-at-length (CAAL) data can improve estimates of constant parametric growth relative to the marginal age data used by Monnahan and Correa (WP 7) (He et al. 2016, Monnahan et al. 2016). It is likely that the benefits of this data type will also improve estimates of time-varying growth in WHAM, but further testing needs to be done to corroborate this expectation and which type of observation model is used for these data could also be a factor.

1.1.3.1 Relevant results from working papers

Li et al. (WP 3) found that an assessment model that did not include missing a source of process variation resulted in biased estimates of the variance parameter of the logistic normal age composition for Atlantic mackerel, and this was especially true when the missing process error was in fishery selectivity. This result suggests that missing a source of process variation in an assessment can be incorrectly absorbed into an estimate of observation variance (Fisch et al., 2021).

1.2 Statistical Reliability

Several papers have been published in recent years that investigate the reliability of state-space assessment models. Jiao et al. (2012) performed a simulation study using Bayesian estimation of alternative process error assumptions (e.g., constant, age-varying, time- and age-varying) on natural mortality for Atlantic weakfish. They found little bias in estimation of SSB, recruitment (R), and fishing mortality (F) whether or not the estimation model assumptions for natural mortality matched those of the operating model, but the worst bias they observed was when a simpler natural mortality assumption was used for the estimation model than the operating model. Miller and Hyun (2018) also performed a simulation study with alternative assumptions about survival process errors and about natural mortality in an application to Acadian redfish and found biased estimation of SSB when a constant natural mortality rate was estimated.

Stock et al. (2021) found, for SNEMA yellowtail flounder, that process errors in both survival and natural mortality provided a better fit than models with process errors in only survival or natural mortality. They also found negligible bias in SSB, F, and R when using models with process errors on survival or natural mortality (specifically) to data simulated without those process errors. Using operating models conditioned on fits to SNEMA yellowtail flounder, Atlantic butterfish, Georges Bank haddock, Icelandic herring, and North Sea cod, Stock and Miller (2021) found the same lack of bias in estimation models with process errors in survival, natural mortality, or selectivity fit to data simulated without these process errors. Results from the simulation study by Fisch et al. (2023) also found increased complexity (more sources of process error) generally does not negatively affect reliable estimation of assessment outputs (i.e., SSB).

1.2.1 Relevant results from working papers

Miller et al. (WP4) also found little evidence of bias in estimation of mean natural mortality for many process error configurations when there was contrast in fishing pressure. However, estimation of natural mortality can cause large differences between the true and estimated SSB (that may be unbiased on average) when there is less contrast in fishing pressure over time and higher observation error. They also found bias in SSB estimation was generally low for estimating models that assumed the correct source of process error when there was lower observation error. When operating models only produced process errors in recruitment estimation models that also included other process errors in survival, natural mortality, selectivity or catchability often provided unbiased estimation of SSB.

The results of the simulation study conducted by Li et al (WP 3) with multiple sources of process error also suggested that incorrectly including additional sources of process error (besides the true process errors) in the assessment models performed similarly to the correct model and generally showed unbiased estimates of management quantities. Both Miller et al. (WP4) and Li et al. (WP3) found that incorrectly excluding a source of process error can produce large biases.

Liljestrand et al (WP 12) performed similar self- and cross- tests of state-space models with various degrees of process error variability in recruitment, survival, and/or selectivity using a fitted Gulf of Maine Haddock WHAM model as the baseline. They found that the models produced unbiased estimates of terminal year SSB and exploitation rate when the estimation model assumed more process variability in recruitment or selectivity than was present. The reverse was untrue- neglecting to account for process variability in survival or selectivity when it was present increased the bias in critical management metrics. However, the variance parameters across most scenarios were estimated with minimal bias.

Monnahan and Correa (WP 7) conducted a simulation study using WHAM with growth included (Correa et al. 2023) where operating models treated initial growth (L1) as constant or as a function of an environmental covariate and found models to estimate accurately when the estimating model assumptions matched the operating model. They also found estimation models that treated initial growth (L1) as an AR1 process when the operating model assumed L1 was constant did not adversely affect model performance.

1.2.1 Self-test simulation studies

Given a fitted assessment model, a self-test simulation study is carried out by simulating a large number of data sets in the same form as that used to fit the original model and then refitting the same model configuration to each simulated data set. Given each of the fitted models, bias in parameter estimation is measured. In traditional applications of statistical catch at age models in stock assessment, there are typically no time-varying parameters other than recruitment. The only random variables are the observations of aggregate catch, indices, and associated age and/or length composition. However, when some parameters are treated as random variables as they are in state-space age structured models, one may either condition on the estimated random effects but simulate the observations or simulate both the random effects and the observations. Simulation studies in the literature that evaluate bias of statistical models that are hierarchical with random effects, seem to invariably simulate both random effects and observations (e.g., Zhu et al. 2010, Aregay et al. 2013, Dorazio et al. 2014, Fasiolo et al. 2016), but, in stock assessment, there is usually an attempt to condition on the perception of how the size of the stock and intensity of fishing has changed over time.

Simulation studies for state-space models by Miller and Hyun (2018), Miller et al. (2018), Stock et al. (2021), Stock and Miller (2021), and Li et al. (WP 3) performed self- and cross-tests for models that did not condition on random effects in keeping with the typical approach in the literature. On the other hand, applications of SAM (Nielsen and Berg 2014) routinely condition on the estimated process errors for self tests. Cadigan (2016) also conditioned on the process errors for self-tests. The WHAM package can be configured to condition on the random effects, and recent self-tests for proposed models within the research track process (e.g., Atlantic cod and black sea bass) have been completed and accepted with this configuration.

Alternative approaches: Marandel et al. (2016) used the unconditional approach for a state-space surplus production model for thorny rays, but they adjusted the exploitation rates for each simulated population to make the pattern in observed catch similar. Given that perception of the evolution of the population and fishing over time is just an estimate, another alternative might be to simulate all parameters and random effects from the estimated variance-covariance matrix assuming normality and the observations given the parameters and random effects. This might be a more appropriate evaluation of bias given the uncertainty in the true latent states being estimated.

Regardless of the method used to conduct simulation self-tests, some bias in parameter estimation should be expected because of the nonlinearity and data limitations. Maximum likelihood estimation is only asymptotically unbiased. Unfortunately, there is no agreed upon standard for the amount of bias that would be deemed unacceptable for management.

1.2.2 Estimating Variance Parameters

An important aspect of state-space estimation models is whether and which variance parameters are estimated. It is also commonly known that accurate estimation and partitioning of observation and process error variances is generally not possible unless observation errors are low relative to process errors even when measurement errors are known (Auger-Méthé et al 2016). Since the earliest investigations of state-space methods in stock assessment, this issue of identifiability of all of the variance parameters has been known (Ludwig and Walters 1981, Walters and Ludwig 1981, Schnute 1994). However, most of these demonstrations of model instability are simple examples with a single latent variable and a single time series of observations whereas in integrated stock assessment there are typically several time series of observations of different types and potentially multiple multivariate latent variables. SAM will estimate by default variance parameters for both process errors and observation errors although observation variances are usually assumed constant over time. Multivariate normal models are the default and most common option for process errors (survival, fishing mortality) and yearly observations at age (catch at age, indices at age) in SAM. The default approach in WHAM is to condition on the variances of aggregate catch and indices provided as input, but to estimate variance parameters for any process errors. Variance parameters for various likelihood options for age composition observations are also estimated by default. The motivation for not estimating observation error variances for aggregate observations is that we usually have good estimates of these for indices from surveys and aggregate catch is usually assumed to be fairly precise or, if discards are included, there are also estimates of uncertainty from observer sampling. Knape (2008) also found that conditioning on external estimates of observation error improved identifiability of state-space models. Variation in sampling effort for indices and age composition over time also complicates estimation of variance parameters for observations. However, the recent application of WHAM for black sea bass estimates a scalar multiplier for the annual observation variances provided for two of the aggregate index time series because the working group found the provided values were implausibly precise.

1.3 Prediction skill

1.3.1 Information criteria

Information criteria have been shown to be useful for comparing relative performance of alternative state-space models. Akaike Information Criterion (AIC) is an estimator of prediction error that penalizes model complexity (Burnham and Anderson 2004). AIC calculated using the marginal likelihood and number of fixed effects (marginal AIC) has been shown to accurately determine best random effects structures in many simulation studies. For example, Celeax and Durand (2008) show AIC performs appropriately for the number of states in hidden Markov models with mixture distributions. Zucchini et al. (2016) also note AIC is appropriate for determining the number of states in hidden Markov models. Miller and Hyun (2018) showed that AIC was completely accurate in distinguishing models with and without random effects on survival for Acadian redfish and also good accuracy for distinguishing between alternative assumptions about natural mortality. Stock et al. (2021) also showed good accuracy of marginal AIC in distinguishing between alternative random effects assumptions for a given aspect of the population dynamics for multiple stocks. In a simulation study evaluating reliability of estimation of mis-reported catch in a state-space assessment model, Perretti et al. (2020) found that AIC selected the estimation model that matched the operating model with high probability when possible. The recent review paper on applying state-space models by Auger-Méthé et al. (2021) notes that no model selection method is perfect and recommends using marginal likelihood-based AIC and WAIC for selecting among state-space models, particularly when the models have the same number of states or random effects. However, the cited simulation studies showed AIC can also be reliable when alternative models have different states and random effects. The ICES WKRFSAM workshop (ICES 2020) also recommends emphasis on out-of-sample prediction (e.g., AIC) rather than goodness of fit for state-space model selection.

Specifically, AIC seems to be helpful in distinguishing between models where parameters are either time-invariant or temporally autocorrelated random effects. For example, models with survey catchability q treated as constant or as an AR1 process with mean, correlation and standard deviation parameters estimated can be compared. Similarly models with and without survival deviation random effects can be compared (Miller and Hyun 2018, Stock and Miller 2021). However, models with annual recruitments treated as fixed or random effects cannot be compared using marginal AIC because of the alternative treatment of each of the annual parameters. When inference on specific random effects is of primary importance AIC calculated using the conditional joint likelihood with an effective number of parameters should be more relevant (Vaida and Blanchard 2005), but calculating the number of effective parameters is not straightforward for state-space models.

On the other hand, Jiao et al. (2012) found poor performance of deviance information criterion (DIC) in selection of alternative natural mortality assumptions for weakfish using Bayesian estimation models in their cross-test simulation study. Similarly, Linton and Bence (2011) found poor performance of DIC in selecting among time-varying selectivity models in a simulation study using Bayesian estimation models.

1.3.1.1 Relevant results from working papers

The simulation studies of Miller et al. (WP 4) where alternative process errors were assumed in operating and estimating models found that the ability of marginal AIC to distinguish the correct model appears to depend on the level of uncertainty in observations and the degree of temporal variation in the model parameters. In cases where observation uncertainty is high and/or temporal variation in the model parameters is low, AIC tends to be more conservative (Type II error) selecting models where the parameter is assumed to be constant. The best performance was with respect to process errors in survival and recruitment where AIC accurately determined the matching assumptions in the estimation model and generally not in the best estimation model when operating models did not include those process errors. However, AIC accuracy for other sources of process error increased with higher variability in the true process error and lower observation error. Similarly, Li et al (WP 3) found that AIC could generally indicate whether an estimating model misspecified process errors (e.g., process errors on survival in the estimating model when the operating model had process errors on selectivity) or a source of process variation was missing. Overly complex models, if converged, exhibited similar AIC to the correct model.

1.3.2 Cross-validation

Cross-validation is the standard general class of statistical methods to measure out-of-sample prediction accuracy. In these methods, models are fit to a subset of data and predictions made for excluded observations. Summary statistics of the error or difference between the observations and predictions are used to measure prediction performance. The most common summary statistics of the prediction errors are root mean squared error (RMSE), mean absolute prediction error (MAPE), mean absolute scaled error (MASE). For some types of statistical models, there are analytic methods to calculate these summary statistics without iteratively fitting models with different subsets of data (e.g., Wood 2006). Furthermore, minimizing cross-validation prediction error is asymptotically equivalent to minimizing AIC for many classes of statistical models (Stone 1977).

Theory and application of cross-validation methods are usually studied for the common scenario where observations are independent and identically distributed. Methods for various dependence structures including temporal, spatial, clustering, and hierarchical correlation also exist (Roberts et al 2017). Integrated assessment models have composition observations which are inherently correlated, but also usually include data sets with different types of observations and with heterogeneity in precision even within each data set. Temporal random effects in state-space integrated assessment models also induce temporal correlation of all of the observations (Auger-Méthé et al 2021).

There has been substantial study of cross-validation methods for time series models which are a special case of state-space models where there is no error in the observations of the time series process. The main two general types of cross-validation methods for models with temporal correlation are “blocked” or “k-fold” and “time series” cross-validation (Hyndman and Koehler 2006, Roberts et al. 2017). In k-fold cross-validation, the time series would be split into k equal intervals and the model is refit k times to data in k-1 intervals (training set), each fit leaving a different interval out of the fitting for predictions (test set). Summary statistics are calculated from the predictions across all k fits. Time series cross validation begins with breaking the time series of observations into training (earlier) and test (later) sets and fitting the model to the training set and comparing predictions from the fitted model during the later test set period with unused observations. The summary statistics are calculated from predictions from rolling fits starting at a minimal length of time to include enough observations that allows the model to be estimated and, moving forward in time refitting the model as the training set gets larger and the test set gets smaller (Hydman and Koehler 2006, Ramos and Oliveira 2016). This is straightforward when there is a single observation each year and all the observations are identically distributed. However, state-space integrated assessment models may include annual fixed effects that are not estimable without an observation. Specifically, annual fishing mortality rates will likely not be estimable without the corresponding aggregate catch observations. Furthermore, there may be numerous observations of different types each year which do not immediately lend themselves to the typical summary statistics used for prediction errors. Lastly, fitting state-space assessment models can be computationally intensive with convergence sensitive to starting values, such that fitting models with incremental increases in the number of observations may be prohibitive in practice.

MASE: Kell et al. (2021) and Carvalho et al (2021) describe different configurations of the MASE statistic for stock assessment and both are different from the description by Hyndman and Koehler (2006) and Ramos and Oliviera (2016). MASE is the ratio of two mean absolute errors for 1) the model of interest (numerator) and 2) a naive model. The naive model may depend on the particular application, but Hyndman and Koehler (2006) and Ramos and Oliviera (2016) use the mean absolute error of the differences between all of the sequential observations in the training set rather than the test set to represent the naive model (a random walk model). The text of Carvalho et al. (2021) suggests these naive model-based prediction errors are from within the training set, but the equation suggests the differences are for the test set. On the other hand, Kell et al (2021), uses a different definition of a naive model in the denominator. In the simulation studies by Li et al (WP 3), they found MASE was generally not useful for model selection, regardless of what horizons and peels were used. They also explored a variant of MASE where the naïve prediction was defined as that from a model without process variances (i.e., a statistical catch-at-age model), but this did not improve the performance.

A limitation of the existing definition of MASE is that it is intended for Gaussian time series models without separate sources of observation and process errors. Aggregate indices often have annual variation in standard errors which is not accounted for in Kell et al (2021) or Carvalho et al (2021), nor do any of the MASE approaches treat multivariate observations such as composition data. Another unsatisfying quality of the MASE statistic as currently applied in stock assessment is that when an index changes very little in the prediction years it is not possible for any model to outperform the naive model.

Using information criteria to compare models requires models to all have the same observations, so cross-validation is particularly appealing when comparing models that use different observations is of interest (Kell et al. 2021, Ramos and Oliviera 2016). However, statistics such as MASE that are used with cross-validation would presumably be comparable only when calculated from the same observations of models being compared.

1.4 Model mis-specification

Although state-space models can accommodate increased realism by allowing parameters in the model to vary temporally and be estimable, we still must ensure that the way we are modeling this temporal variability (as a stochastic process and/or as a function of environmental covariates) on the population is consistent with the data used to fit the model. For example, we may allow recruitment to vary over time, but perhaps we mis-specify the nature of any autocorrelation in recruitment. In contemporary stock assessment there are two primary sources of information that are inspected to evaluate model mis-specification: residuals and retrospective patterns.

1.4.1 Residuals

Inspection of residuals for patterns over time and age and cohort for composition data is a common approach to assess model mis-specification. For correctly specified models, quantile and Pearson residuals should be standard normal in distribution. However, most formulations for these residuals assume all observations are independent, but not necessarily identically distributed. Current practice is to report Pearson residuals for all data types. Pearson residuals are appropriate for aggregate catch and index observations when traditional statistical catch at age models are used. However, Pearson residuals for age, length and age-at-length composition observations violate the assumptions of independence and being standard normal distributed because of the inherent correlation of the vector of observations at age, length or age at length (Trijoulet et al. 2023). Furthermore, when random effects are used to model temporal variation in the dynamics of the population, all observations become correlated. The current best practice for assessment models with correlated observations is to report and analyze one-step-ahead (OSA) residuals (Zucchini et al. 2016, Thygesen et al. 2017, Auger-Méthé et al. 2021, Thorson et al. 2023, Trijoulet et al. 2023). One step ahead residuals can be calculated using the TMB package and both WHAM and SAM can report these for all observations.

To use the TMB functionality, data must not be transformed in the C++ code of the model. For example multinomial observations must be provided as frequencies to the compiled model rather than multiplying sample size and the observed proportion in the compiled model. The OSA residuals can be analyzed similarly to traditional analysis of residuals. However, when patterns in OSA residuals are apparent for state-space models, determining the mis-specification causing the patterns is challenging. Investigations of the relationships of mis-specification and OSA residual patterns should be a high priority.

WHAM currently generates various plots of OSA residuals. For aggregate observations, there are q-q plots, histograms, and plots of residuals vs. year and vs. predicted value. For composition residuals there are also q-q plots, histograms, and plots of residuals vs., year, age, cohort, and predicted value.

Ramos and Oliveira (2016) used statistics (e.g., RMSE) of one-step-ahead predictions/residuals as a model selection criterion for state-space and ARIMA models. They noted it was useful because AIC could not be used to compare alternative ARIMA models where the observations were different due to transformations. This situation is not as relevant in assessment models except perhaps for alternative treatment of composition data. However, Albertsen et al. (2017) demonstrated the use of change in variables for the distributions so that comparable AICs can be calculated.

1.4.1.1 Relevant results from working papers

In the simulation study by Li et al. (WP3), they found that the mean of the OSA residuals was unlikely to be useful to determine correct process error configuration, but that relatively large variances of the OSA residuals (SD > 1) were indicative of a missing source of process error in some cases. Normality tests (Anderson-Darling and Kolmogorov-Smirnov) were also able to identify a missing source of process variance in some cases. Neither the variance of the residuals nor the normality test was able to choose the correct model among EMs of similar complexity.

1.4.2 Retrospective patterns

Retrospective patterns are defined by systematic changes in terminal year estimates as new data are added to the assessment model (Mohn 1999). These patterns are understood to arise when there is a disparity between the assessment model configuration and the true data generating mechanism. The most common retrospective pattern indicated by the Mohn's rho diagnostic is a positive rho for SSB and a negative rho for F. This pattern indicates that SSB estimates in each recent year tend to decline as years of data are added to the model and vice versa for F estimates. The mechanisms that can explain such a pattern include differences between true and assumed natural mortality or systematic differences between reported and true catch.

Allowing temporal variation in aspects of the assessment model that would otherwise be constant has been shown repeatedly to reduce retrospective patterns. Current practice is to estimate assessment models with annual data sequentially removed from the estimation model (peels) and report the Mohn's rho statistic for fishing mortality, spawning biomass and recruitment. Large absolute values of Mohn's rho are undesirable. The number of peels to use when calculating Mohn's rho is subjective, but seven are typically used in the Northeast US. It has also been standard practice to "retro-adjust" terminal year estimates of abundance at age, F, and SSB for short-term projections to make catch advice which was shown by Legault et al. (2023) to provide management advice generally as well as simpler empirical approaches.

Martell and Stewart (2014) found including time-varying selectivity to eliminate retrospective patterns in an assessment of Pacific halibut. Linton and Bence (2011) found Mohn's rho to be useful in model selection to determine models with less biased estimation of model output (e.g., SSB).

Like all summary statistics, Mohn's rho is an estimator with variance. The uncertainty of Mohn's rho is difficult to estimate and therefore not typically reported in assessments. Bootstrap methods can provide accurate estimates of uncertainty and confidence intervals but it is time-consuming for assessment models even without random effects (Miller and Legault 2017). Brevik et al. (2023) propose a model-based approach involving simulating the most recent years of data, but their work showed typically poor power to detect a Mohn's rho different from zero (Type II error). Recent work in the ICES Methods Group has developed an approach to directly estimate the standard deviation of Mohn's rho in state-space models (see section 11.19 in https://www.ices.dk/about-ICES/Documents/Resolutions/EG%20files/MGWG%202022%20abstract%20of%20presentations.pdf).

1.4.2.1 Relevant results from working papers

In the simulation study by Miller et al. (WP4), they found retrospective patterns were generally weak for all estimation models regardless of the true source of process error, but they can be expected for recruitment even for the correct process error assumptions when observation error is high. When models did exhibit some retrospective pattern, estimating the mean natural mortality rate tended to remove it. Li et al. (WP 3) found adding random effects generally reduced retrospective patterns. Strong retrospective patterns (|Mohn's ρ| > 0.2) could emerge due to the above-mentioned two types of model misspecification. Overly complex models, if converged, exhibited similar retrospective patterns to the correct model.

1.4.3 Posterior predictive checks

Gelman et al. (2004), Conn et al. (2018) and Auger-Méthé et al (2021) recommend this as a diagnostic for Bayesian models including hierarchical models. Hobbs et al. (2015) uses this diagnostic for a state-space model of infection in a bison population. Conn et al. (2018) showed that the power of the approach described by Gelman et al. (2004) is poor (tends to show the model is good: Type II errors). Conn et al (2018) also showed that the method should only simulate 1 posterior set of random effects then simulate *N* sets of observations from that. For a TMB state-space model, we have posterior empirical Bayesian estimates of random effects. Steps to compute this diagnostic are: 1) simulate the random effects from the posterior once (as in Thygesen et al. 2017) 2) simulate many sets of observations 3) calculate summary statistics for the true observed data. 4) calculate summary statistics for each simulated data set. 5) compare the summary statistics in 3) (i.e., as a quantile) to the distribution of the simulated summary statistics. SAM uses the methodology of Thygesen et al. (2017) to perform posterior check of random effects.

1.6 Other Diagnostics

1.6.1 Convergence

When fitting assessment models one of the first issues that can arise is lack of convergence when attempting to optimize the model (i.e., maximize the likelihood). Convergence has been measured several ways including the maximum absolute gradient component at the optimized objective function, and whether the hessian at the optimum can be inverted (matrix is positive definite), but convergence can also be measured more simply by whether the optimization even successfully completes. Checking parameters for bounds or large standard error estimates is recommended once the convergence criteria (e.g. a low final gradient or invertible Hessian) have been met. Additional convergence diagnostics, such as jittering model parameters, can also be performed to ensure that the model reaches a global solution rather than a local minimum (Carvalho et al. 2021). Jittering analysis may also be useful when the model has not yet converged. In a recent study (Fisch et al. 2023), they jittered the initial values of model parameters from unconverged models repeatedly (e.g. 5 times) before concluding that the model had not converged.

For converged models that estimate autocorrelation parameters across ages and/or years for random effects, it is prudent to examine the sign of these estimates. If estimates are negative there should be some hypothetical mechanism to explain why adjacent ages or years would be negatively correlated.

1.6.1.1 Relevant results from working papers

In the simulation study completed by Miller et al. (WP 4), they found alternative measures of convergence performed differently. Invertible hessians and resulting standard error estimation was possible when criteria based on the gradient of the optimized log-likelihood with respect to the fixed effects parameters failed (Carvalho et al. 2021). Using hessian-based convergence, probability of convergence was best for models that assumed the correct source of process error, assumed natural mortality was known, and did not assume stock-recruit relationships.

In the simulation studies by Miller et al. (WP 5) focused on estimation of environmental covariate effects on natural mortality, they found convergence (based on whether the hessian could be inverted) was generally best when operating models assumed process errors in recruitment and survival, constant fishing rate, greater contrast in the true environmental covariate, and lower uncertainty in corresponding observations. Reliable convergence also occurred when estimating models used the correct assumption about process errors and there was a step-change in fishing, but this also required lower uncertainty in index and age composition observations. Estimating models with process errors on recruitment and survival were unlikely to converge when the process errors in the operating model did not match; whereas estimating models with process errors in recruitment and natural mortality converged for operating models without this match in certain cases. Probability of convergence generally decreased when the mean natural mortality rate parameter was estimated.

Li et al. (WP 3) in their simulation studies, found that a high convergence rate did not guarantee a more accurate assessment model. Specifically, correctly specified estimating models and overly simplified estimating models were both likely to converge. An estimating model with a wrong process error could also converge, likely by using its free parameters to explain the true varying process in the operating model. For example, allowing numbers-at-age deviations to vary in the estimating model seemed to facilitate the model convergence when selectivity or natural mortality was the true varying process in the operating model.

1.6.2 Likelihood profiling

Likelihood profiling can be used to estimate confidence intervals or just inspect the degree of curvature of the log-likelihood surface with respect to the parameter being profiled. It is also common to inspect the change in the log-likelihood components for each type of observations used to fit the model (e.g., aggregate indices, catch, age compositions). It is common to do profiles for natural mortality and unfished recruitment (for models that use this parameter) (Carvalho et al. 2021). For state-space models that are estimated by maximizing the marginal log-likelihood, calculating marginal-log-likelihood components for each data type is not straightforward. Perreault and Cadigan (2021) propose and demonstrate a method to provide profile components by calculating the marginal log-likelihood with and without the data component included, but there is no guarantee that the inner optimization of any random effects without all of the data will converge. These authors also note that an individual joint negative log likelihood data component may indicate a better fit at a lower M, but requires very large random effects to do so.

1.7 Recommendations for selecting preferred state space assessment models

1. **Treat recruitment as random effects so that variance and correlation parameters can be estimated, but use model selection methods to determine an appropriate time series model for the latent annual recruitments to ensure reliable projections.**
2. **When considering which parameters to treat as time and age-varying, err on the side of using overly complex models. If these models estimate no variability in particular process errors, then those process errors can safely be removed for parsimony and better convergence properties. However, caution is warranted with process error on natural mortality as it has been shown to result in biased estimation of model output for management in some scenarios and the resulting natural mortality estimates have direct consequences for management reference points.**
3. **When non-negligible mis-reporting of catch is plausible, estimation of catch process errors should be considered, and estimated errors inspected for bias (i.e. can help reveal under-reporting).**
4. **When reliable external estimates of observation error variance are available treat them as known in the assessment model, particularly when they are low relative to process errors. When measurement error variance is large, self test simulations should be used to ensure the model is reliable.**
5. **Perform posterior check of all random effects as described by Thygesen et al (2017) for evidence of model mis-specification**
6. **When using MASE with time-series cross-validation, we recommend using the denominator as described by Hyndman and Koehler (2006). When there are multiple indices and composition observations each year, rolling fits should not incrementally include each type of observation in a given year, because they are correlated due to the autoregressive process errors. We also recommend a generalization of MASE that uses randomized quantile prediction errors as described by Thygesen et al (2017) for one-step-ahead residuals. The generalization of one-step-ahead residuals for multivariate observations by Trijoulet et al (2022) might also be applied to include composition observations in MASE. Note that one-step-ahead residuals do not require refitting the model iteratively and therefore can also include residuals for aggregate catch observations. These observations cannot be included in MASE because, in general, they cannot be excluded for model fitting when annual fishing mortality parameters are estimated as fixed effects.**

7. **Use a broad suite of metrics and diagnostic tools to evaluate relative performance of alternative models. Statistical reliability and AIC as a model selection tool are better when there is contrast in fishing pressure, stock size and process errors over time and more precise index and age composition observations are available.**

TOR 2. Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.

2.0 Introduction

Here we first review previous work on estimating stock-recruit functions using state-space methods. We also review relevant results from working papers completed by the WG, and end the section with recommendations for estimating stock-recruit relationships within state-space assessment models.

2.1 Background

Stock-recruitment relationships are notoriously difficult to estimate due to multiple confounding sources of variability. Subbey et al. (2014) suggest that many in the field think of reliable stock-recruitment estimation as an unreachable goal. One reason for this difficulty is that observations of recruitment occur after a myriad of density-dependent and density-independent processes have occurred, resulting in a very high degree of inherent process variability. The final product of multiple density-dependent processes can appear relatively flat, further contributing to the challenge of model identification (Brooks et al. 2018). Recruitment and spawning stock biomass are also difficult to observe resulting in a high degree of measurement error.

State-space modeling is an appropriate statistical approach to inference when both measurement and process error are present, including stock-recruit relationships. There have been two general alternative approaches to configuring process errors in state-space models of stock-recruitment relationships. In the first, the density-independent and density-dependent stock-recruit parameters are defined as the latent processes with process error. In the second, recruitment is the latent process with the mean defined by the stock-recruit relationship.

2.1.1 Process errors in the stock-recruit parameters

Peterman et al. (2000) used a state-space approach to estimate time-varying stock recruitment parameters as time-dependent process errors. The method has since been termed the Peterman Productivity Methods (PPM; Silvar-Viladomiu et al. 2022), although the initial credit for the approach should be given to Walters and Hilborn (1976). Peterman et al. (2000) focused the Kalman filter approach with the density-independent slope at the origin of the stock recruitment function as the time-varying latent process.

Dorner et al. (2008) applied the PPM to 120 Pacific salmon stocks to estimate broad patterns in estimated productivity time series. Minto et al. (2014) extended the PPM to the multivariate case where parameter variation was modeled at a multivariate random walk and estimated the covariance matrix of the multivariate Gaussian errors in an application to. recruitment and spawning stock biomass output from stock assessments for cod stocks across the North Atlantic using. Britten et al. (2015) used the PPM method to estimate process error time series for hundreds of globally distributed stocks using recruitment and spawning stock biomass estimates from a global stock assessment database as the ‘data’ in this case.

2.1.2 Process errors in the recruitment

de Valpine and Hastings (2002) investigated the Beverton-Holt and Ricker recruitment models as population dynamics models and found that state-space models led to lower bias and often lower variance estimates than least squares estimators that ignore either process noise or observation error. The paper also showed poor model selection performance for the correct stock-recruit model assumption. However, the simulations and estimation models assumed independent and homoscedastic observations of SSB and recruitment whereas these estimates from stock assessment models are correlated and heteroscedastic. Maunder and Deriso (2003) compared age-structured models with alternative approaches to estimating recruitment including penalized likelihood, marginal likelihood and Bayesian and found the latter two approaches provided better estimation of annual recruitments and the variance of the annual recruitments. However, simulation and estimation models did not assume any stock-recruit relationship.

State-space assessment models such as WHAM and SAM allow for process errors in both recruitment and spawning stock biomass and can estimate stock-recruitment relationships within the stock assessment model. As previously described, WHAM can also estimate alternative effects of environmental covariates on recruitment and stock-recruit parameters, but it does not allow the stock-recruit parameters to be modeled as latent temporally varying random effects.

2.1.3 Relevant results from working papers

The Britten et al. study (WP 1) simulated 64 OMs with a Beverton-Holt (BH) stock recruitment relationship (SRR) without an environmental effect and fit two EMs to those simulations that did not include an environmental effect - one with mean recruitment and one with a BH SRR. The authors found poor identifiability of the BH SRR based on AIC where mean recruitment was favored in the majority of cases despite an underlying BH SRR. The situation improved (correct model identified more often) with high contrast in fishing history and low recruitment standard deviation. The authors also found that both SRR and mean recruitment EMs generally yielded unbiased assessment quantities (estimated R, SSB, and F); however, they did find minor effects of how OM factors contributed to relative error. Specifically, the effects of recruitment standard deviation and fishing history were anticorrelated among EMs with and without an SR relationship. High recruitment standard deviation and MSY fishing history generally led to more-positive relative bias for EMs without an SRR and led to more negative relative bias in EMs. This was more pronounced when assessing bias in the last year of the assessment.

The structure of the simulation study design by Miller et al. (WP 4) and results are summarized in Appendix 3. The authors found probability of convergence, as measured by an invertible hessian at the optimized log-likelihood, was best for models that assumed the correct source of process error, assumed M was known, and did not assume stock-recruit relationships. The study results also indicate AIC more accurately determined a Beverton-Holt stock recruit relationship rather than the null model without a S-R relationship when there was low variability in recruitment, low variability in survival random effects, and higher variation in spawning biomass over the time series. Similarly, reliable estimation of Beverton-Holt stock-recruit relationship parameters only appears possible in ideal situations with lower observation errors in age composition and indices, lower variability in recruitment process errors and large contrast in spawning biomass over time. For combinations of operating and estimating models where there was bias in natural mortality due to high observation error, estimating the stock-recruit relationship seemed to remove the bias.

2.2 Recommendation on estimating stock-recruit relationships in state-space assessment models

1. Consider the level of information in the stock assessment data for the stock-recruit relationship. Answering yes to more of these questions increases the likelihood of accurately determining the stock recruit relationship and reliability in the parameter estimates.
    a. ***Is the time series sufficiently long?*** The longer the time series, the greater the number of years with recruitment and spawning stock biomass estimates and the greater the likelihood of there being sufficient information in the stock assessment data to accurately estimate the stock-recruit relationship. Conn et al. (2010) also found this to be important for non-state-space models.
    b. ***Is there evidence of good contrast in spawning stock biomass over time?*** Reliable estimation of the stock-recruit relationship parameters requires spawning stock biomass over the range of the curve, and away from the asymptote for the Beverton-Holt relationship. Large range in spawning stock biomass was found by Miller et al. (WP 4) and Britten et al. (WP 1) to improve inferences on stock-recruit relationships in state-space models. Magnusson and

    Hilborn (2007), Conn et al. (2010), and Lee et al. (2012) also found contrast in SSB important for reliable estimation of stock-recruit parameters in assessment models without random effects estimated by maximum marginal likelihood.

    c. ***Are index and age composition observations relatively precise?*** Miller et al. (WP 4) found high observation error will make it more difficult to detect a stock-recruit relationship and less reliable estimation of stock-recruit parameters. Conn et al. (2010) also found this to be important for non-state-space models.
    d. ***Is variation in recruitment residuals (sigma-R) relatively low?*** Miller et al. (WP 4) and Britten et al. (WP 1) found better detection of the Beverton-Holt stock-recruit relationship and reliable parameter estimation with lower variability in residual recruitment.
2. Estimate the stock-recruit relationship simultaneously and internal to the state-space stock assessment model.
    a. This will allow variation in uncertainty in recruitment and spawning stock biomass, and correlation among these assessment model outputs to be properly propagated into the estimation of the stock recruit relationship.
    b. If done externally, use methods that would treat recruitment and spawning stock biomass in a multivariate fashion to account for uncertainty and correlation among the annual values.
3. Self-tests as described in TOR 1 would be prudent to confirm reliability of stock-recruit parameter estimates and biological reference points derived from them.
4. Consider alternative autocorrelation models for residual recruitment. This will be important primarily in defining how recruitment is predicted in short-term projections.

TOR 3. Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.

3.1 Introduction

We review previous work and work conducted by the working group that is relevant to this Term of Reference. We make recommendations on the best diagnostics to use when determining whether or not to include climate effects and situations when we can reliably detect and estimate environmental effects and accurately estimate assessment output (SSB). The software developed to conduct the simulation studies carried out by the working group provides a framework to explore the inclusion of environmental effects in future assessments.

3.2 Background

Modern stock assessments integrate different sources of data to produce estimates of important population dynamics processes and trends over time. In recent years, it is well documented that ocean conditions are changing, affecting species distribution, abundance and productivity (Nye et al., 2009; Pinsky et al., 2013). All of which have direct implications for population assessments and management. Failure to account for these changes can lead to inaccurate understanding of population dynamics and trends (Mazur et al., 2023). Despite this, the majority of stock assessment and management do not directly account for changing ocean conditions (Pepin et al., 2022 more below). Thus, a research priority is figuring out how changing environmental conditions can be incorporated into assessments and management sufficiently well.

Environmental covariates can be incorporated into stock assessments in a variety of different ways. Ecosystem indicators can be used to form hypotheses and provide context to assessment results. Analyses can be conducted outside of the assessment with results included as data inputs (e.g., catch rate standardization that produces an index). Information on environmental conditions can also be used to inform prior distributions for parameters within the assessment (Romakkaniemi 2015). Modern assessment platforms like WHAM can also directly fit to environmental time series and estimate parameters relating environmental data to ecological processes (Stock and Miller, 2021).

Traditionally, most work has been given to exploring environmental effects on recruitment, primarily because for many species there is substantial temporal variation in recruitment and fisheries scientists have a long history of debating whether recruitment success is better attributed to population size or changes in the environment (Maunder and Thorson, 2019). A series of environmental hypotheses have been developed within fisheries science and fisheries oceanography describing potential recruitment-environment links (Hjort 1914; Hare 2014). These include, among others, the aberrant drift hypothesis (Hjort 1914), critical period hypothesis (Hjort 1914), match-mismatch hypothesis (Cushing 1990), and stable ocean hypothesis (Lasker 1981). While these hypotheses contain different (although sometimes overlapping) mechanisms, all invoke the external environment as a driver of recruitment success. A synthesis study from Szuwalski et al. (2015) that analyzed 224 stocks found that recruitment was often uncorrelated with spawning stock biomass, suggesting that environmental factors may be a better indicator of recruitment success. Additionally, assessments have found that directly fitting to climate indices can reduce uncertainty in recruitment estimates (Sculley et al. 2018) and lead to reduced retrospective patterns and improved recruitment predictions (du Pontavice et al 2022).

Assessments have also explored incorporating ecosystem covariates on: growth, catchability and natural mortality. Lee et al. (2018) demonstrated that including climate indices as modulates of growth increased precision and reduced bias in assessment output and found that including indices when growth was constant did not result in biased estimation. ICCAT assessments that have multiple area-specific catch rate indices have used climate indices to account for distribution shifts by incorporating water temperature to account for time-varying catchability (Hansell et al., 2020; Schirripa et al. 2017). Using climate indices as modulates of catchability improved parameter estimates and reduced residual patterns (Hansell et al., 2020). Predator indices have also been fit to in a variety of assessments to account for changes in natural mortality and including these indices can lead to more accurate estimates of stock biomass (Hollowed et al., 2000).

It is more common for stock assessments to use time-varying parameters without a mechanistic link, referred to as the ‘implicit approach’ (Punt et al. 2014). These can help to improve fits to data without the need to model explicit effects of covariate. Improved fits can then be attributed to multiple mechanisms (e.g., changes in management, movement, or the environment) instead of a specific hypothesis. Incorporating generic time-varying parameters can make models more realistic as well as reduce bias in parameter estimates and lead to more accurate perception of stock trends (Deroba and Schueller 2013; Trijoulet et al. 2020). By describing time-varying parameters as time-dependent random effects, state-space models provide a well-grounded and flexible mathematical framework for including time-varying parameters that allow for different assumptions about the underlying parameter dynamics, including autocorrelation. Including random effects also adds flexibility to the model that can account for variance in ways not reflective of the underlying processes.

Despite the desire to include ecosystem effects in stock assessment there are relatively few instances of environmental indices being directly included in assessments (Skern-Mauritzen et al. 2016, Peppin et al. 2022; Marshall et al. 2019). Further, most of the assessments that do directly fit to ecosystem effects are in a reduced state (Marshall et al. 2019), which is most likely the result of strong contrast in the data required to estimate such parameters. The primary reasons more ecosystem effects have not been included in assessments are: 1) their relationship with population dynamics processes are difficult to observe; 2) the relationship often changes over time; and 3) the typical scale of data collection between fisheries data does not match the scale environmental variables are measured (Maunder and Thorson, 2019; Crone et al. 2017). Stock assessments are also intended to be simplified representations of a population and adding unnecessary complexity (e.g., climate indices) when not well understood can lead to degraded model performance. In contrast, failure to account for a climate index when one is present can also lead to poor assessment performance (Mazur et al. 2023).

3.3 Methods for including explicit environmental effects

Schirripa et al. (2009) described “model” and “data” approaches to including covariate effects on recruitment. Miller et al. (2016) and Miller et al. (2018) showed how mechanistic and implicit approaches (Punt et al. 2014) and data and model approaches (Schirripa et al. 2009) can, and perhaps should be, combined for making inferences on mechanistic effects of specific covariates on recruitment and growth, respectively. They compared models that assume stochastic variation in the parameter, but with and without effects of a covariate. When a covariate effect provides improved model performance, information criteria and residual variation in the parameter it effects will be lower. Miller et al. (2018) also showed that only comparing models that assume no stochastic variation in growth, but with and without temperature effects, leads to dramatically different perceptions of the temperature effects on growth and AIC values that are very poor relative to any models that also include stochastic variation in growth. Maunder and Watters (2003) similarly demonstrated the need to include stochastic variation in recruitment when evaluating covariate effects. It is standard to begin with a model with annual variation in recruitment when investigating covariate effects, but not other demographic parameters. An open question is the appropriate correlation structure of the random effects for the null model (and that with the covariate effect added).

A common practical issue with analyses of effects of environmental covariates within assessment models is how to deal with periods of time where covariates are not available. Maunder and Deriso (2010) examined alternative approaches for dealing with missing observations and found treating them as a random effect was preferred. The state-space approach used for environmental covariates by Miller et al. (2016) and Miller et al. (2018) treats covariates as a latent time series process with random effects so that the covariate can be predicted where the covariate observation is missing, but the predictions are still informed most by observations that are near in time to the missing periods.

3.4 Accounting for covariate observation error

State space models also provide a framework for incorporating measurement error in environmental covariates. As Walters and Ludwig (1981) point out, measurement error in spawning stock biomass can lead to weak correlations between recruitment and spawning stock biomass, leading to erroneous conclusions that there is no relationship between the two when there is (i.e. Type II error). The same can happen for environmental relationships when measurement error for an environmental variable is high. By explicitly allowing for measurement error in the environmental variables (e.g. Miller et al. 2016) state-space models may provide more accuracy in detecting relationships when fitting to noisy environmental data or, if there is uncertainty, how the environment data should be processed for analysis. This is also relevant for cases when fitting to climate model output instead of direct measurements (e.g. Xu et al. 2018, du Pontavice et al. 2022), when there can be errors in the representation of environmental processes within the model.

3.5 Methods for detecting effects of environmental covariates

In practice, assessment scientists must first evaluate whether there is evidence for an effect of an environmental covariate on a particular parameter in the assessment model if there is evidence of temporal variation in the parameter. De Oliveira and Butterworth (2005) describe typical steps in evaluating whether the effect of an environmental covariate on recruitment is evaluated externally to an assessment model and they include this in their simulation study. Haltuch and Punt (2011) compared methods to evaluate effects of covariates on recruitment that included external or internal to the assessment model and found that methods internal to the assessment model exhibited high probability of rejecting the null model for effects of environmental variability on recruitment when the null is true (spurious correlation, Type I error). On the other hand, Maunder and Watters (2003) found low Type I error for internal estimation of environmental effects on recruitment in their simulations.

Maunder and Watters (2003) recommended likelihood ratio tests to determine whether or not environmental variables should be included. While this a valid approach assessment models have evolved since 2003 and current best practices do not provide recommendations for state space models that can include a wide variety of environmental variables and random effects in the same framework (Carvalho et al. 2021).

However, evaluating effects of environmental covariates is better performed within the assessment model to account for uncertainty and correlation of all parameters in the model that would also affect the particular demographic aspect of interest (e.g., Miller et al. 2016 and Miller et al. 2018). The improvement could be measured using the methods (e.g., AIC) reviewed in TOR1.

3.6 Accounting for environmental variation on biological reference points and management advice

For many stocks of New England groundfish, projections have a history of being positively biased, resulting in catch limits being set too high (Wiedenmann and Jensen 2018). The projection bias was largely the result of strong, unaccounted for retrospective patterns, but also from below-expected recruitment and decreasing size-at-age for some stocks (Brooks and Legault 2016, Wiedenmann and Jensen 2018). Recent work suggests that the assessment bias may be environmentally-driven (Kerr et al. 2022). Therefore, incorporating process errors in model parameters or explicit effects of environmental covariates on them could potentially improve management advice in the region if it reduces retrospective patterns or improves the estimates of recruitment or post-recruit productivity components used in the projections.

The standard practice in the NEUS is to use SPR-based reference points with estimates of productivity under prevailing conditions. Traditionally, weight, maturity, selectivity, and natural mortality at age are averaged over recent years and used to calculate equilibrium spawning-biomass-per-recruit and yield-per-recruit or in long projections with constant fishing mortality rates. Recruitment is treated stochastically in these projections, but there is variation among stock assessments in how this stochasticity is generated.

In ICES, the approach for calculating reference points similarly emphasizes prevailing conditions in accounting for temporal variation in productivity. The management process involves re-estimating reference points frequently, at most every 5 years, to avoid large changes in reference points and stock status, and provide stability in catch advice (ICES 2021, 2022).

The traditional tools that assessments have used for reference points (ASAP and AGEPRO) do not allow explicit covariate effects on reference points. Using WHAM allows internal estimation of MSY- and SPR-based reference points so that uncertainty in parameters being estimated in the model can be propagated into estimates of uncertainty for the reference points. Effects of explicit environmental covariates on recruitment and natural mortality can be considered, and the effects on these productivity components will also consequently affect reference points.

3.7 Summary of relevant results from simulation studies completed by the working group

3.7.1 Britten et al. (WP 1)

In the Britten et al. study (WP 1), the authors found high convergence rates of the estimating models (>95%) except in cases with low recruitment variability and constant fishing history. In >20% of these cases, the model failed to fit and caused R to abort. The authors look forward to follow-up work to better understand the cause.

In general there was low identifiability of an underlying stock-recruit model (33% correct); although rates of identification increased to above 50% with high contrast in fishing history and low recruitment variability. Identifiability was also low for the correct functional form of the environmental covariate relationship (29%). Rates of correct identification depended on the strength of the environmental effect, where higher effect sizes led to higher identification rates, and lower recruitment standard deviations led to high identification rates. Highest identification rates occurred for mean recruitment with no environmental covariate, exceeding 80%. Correct identification of both the stock-recruit relationship and environmental covariate relationship was very poor (10% overall), but was most successful with high contrast in fishing, low recruitment variability, and the null environmental covariate relationship.

All estimating models tended to fit the data similarly well, resulting in the average difference in AIC of 2.8 between the best and second-best fitting models (1.7 if recruitment variability was high), indicating support for many of the estimating models applied to the same data simulated from a given operating model.

Relative error was summarized over all years, the last 10 years, and the final year. Estimates were generally median unbiased and didn't differ among estimating models (the misspecified estimating models did as well as the correct estimating model). Higher recruitment variance led to wider range of relative errors in model estimates of recruitment, but less so for spawning biomass and fishing mortality. In generalized linear (mixed) model-based effect size estimation, recruitment relative error depended most strongly on recruitment standard deviation and observation error where lower variance led to lower relative error, but the effects were generally small. The effect of recruitment random effect correlation increased for the final year where high correlation led to more positive relative bias. Spawning biomass followed the same general pattern as recruitment. Relative error for fishing mortality decreased with a high recruitment standard deviation.

Parameter bias showed interesting patterns, with recruitment standard deviation and fishing history playing the largest role, with exceptions. Parameters were median unbiased except for random effects process parameters. Latent environmental covariate standard deviation had positive relative bias when simulated correlation of latent covariates was high. Recruitment random effects parameters had bias that traded off - relative bias for recruitment standard deviation was positive when recruitment correlation was high, whereas recruitment correlation relative bias was negative when simulated recruitment correlation was low. Environmental effect size was the most poorly estimated parameter with the quantile range of relative error exceeding 20.

Median Mohn’s rho was approximately zero, but the range of values increased when recruitment variance was high and there was low observation error. For spawning biomass and fishing mortality, there was a slight negative bias in median Mohn’s rho for spawning biomass and slight positive bias in median Mohn’s rho for fishing mortality. The retrospective pattern on the recruitment random effects generally had negative bias and a much wider range of values compared to recruitment, spawning biomass, and fishing mortality. The random effect retrospective was not lower for the correctly specified estimating model.

There was no discernable difference in the projections relative to two assumptions about the environmental covariate in the future (continue the estimated process or project at a value calculated as the mean of the last 5 years).

3.7.2 Hart and Hansell (WP 2)

Hart and Hansell (WP 2) evaluates state-space assessment performance across a range of observation uncertainties, fishing histories, and environmental effect sizes for models with the following catchability assumptions: 1) time-varying, 2) environmentally-driven, 3) both time-varying and environmentally-driven, and 4) constant (status quo) catchability. In total, 384 operating models were used for the different simulations.

Median convergence rates were near one for status quo and models with only an environmental covariate on fall survey catchability, but of these, status quo models had more outliers with lower convergence rates. In contrast, models with catchability random effects (alone or with an environmental covariate) tended to have convergence rates below one. EMs with catchability random effects generally converged at similar or higher rates than those with both catchability covariates and random effects and convergence rates tended to increase with OM environmental effect size. Convergence rates dropped below 1 for all EMs when the OM had a small environmental process error and large environmental observation errors often with lower convergence rates. These patterns remained fairly consistent across fishing histories and seasonal misspecifications.

All EMs had similar median recovery of reference points (Fmsy, MSY, SSBmsy), but status quo models more frequently overestimated these reference points as environmental effect size increased to intermediate and high levels. Similar patterns of overestimation were also observed under these conditions for all models without catchability random effects when one season was misspecified and for all EMs when both seasons were misspecified.

AIC always selected status quo models when the environmental effect size was 0, but selection differed across seasonal misspecifications at larger effect sizes. When both seasons were correctly specified, AIC most often selected for the correctly specified EM with environmental covariates on the fall survey catchability, followed by status quo models. At low environmental effect size and one seasonal misspecification, AIC selected for status quo models, but under larger effect sizes, EMs with only catchability random effects were selected. Finally, AIC always selected status quo models most frequently when both seasons were misspecified, followed by EMs with only an environmental covariate, but selection of EMs with random effects (with or without covariates) increased with environmental effect size.

WHAM models with catchability random effects (with or without environmental covariates) had marked trade-offs between model reliability, selection, and performance consistency. When seasonal drivers of catchability were reasonably informed (one or no seasons misspecified) then median model performance was typically similar to the correctly specified model. However, these models often had lower precision for some estimated values (e.g. selectivity- and catch-at-age) and consistently lower convergence rates. This suggests that adding catchability random effects to a season with no climate-driven change does not degrade model performance on average, but relying on catchability random effects to account for time-varying change can increase variability in estimates of some key parameters and reduce convergence rates. Models with both random effects and environmental covariates tended to have the lowest median convergence rates of all EMs, suggesting that overfitting reduces model reliability. However, they also recovered environmental effect size with more precision than models with only a covariate (Figure 4), indicating that some of the variance in the environmental covariate has been partitioned to the catchability random effect.

Estimation model performance varied across seasonal misspecifications in a predictable way and highlighted broad challenges that WHAM models are likely to encounter when environmental drivers are misrepresented. Status quo models consistently produced biased estimates of performance metrics when index observation error was small, environmental process error and effect sizes were large, but similar performance issues were not observed for other EMs under these conditions until seasonal misspecifications were introduced. These results suggest that status quo models are most likely to struggle to describe systems with strong but noisy environmental drivers of catchability for well observed indices, but that other EMs also struggle to describe these systems when provided incorrect assumptions about impact seasonality.

Overall, our results suggest that incorporating random effects or environmental covariates into stock assessment models improves their ability to characterize time-varying catchability when there is a moderate or strong environmental trend. When environmental relationships are well understood and seasonally correct, models with only an environmental covariate are able to account for time varying catchability. However, if the environmental trend is not well understood or is seasonally misspecified, random effects are more effective at accounting for time varying catchability.

3.7.3 Miller et al. (WP 5)

The structure of the simulation study design by Miller et al. (WP 5) and results are summarized in Appendix 3. The authors found convergence of all estimation models was generally best when operating models assumed process errors in recruitment and survival, constant fishing rate, greater contrast in the true environmental covariate, and lower uncertainty in corresponding observations. Reliable convergence of all estimating models also occurred with the same process errors in the operating model and a step-change in fishing, but this also required lower uncertainty in index and age composition observations. Estimating models with process errors on recruitment and survival were unlikely to converge when the process errors in the operating model did not match whereas estimating models with process errors in recruitment and natural mortality converged for operating models without this match in certain cases. Probability of convergence generally decreased when the mean natural mortality rate parameter was estimated.

Whether the mean natural mortality rate parameter was estimated or not, the best accuracy of marginal AIC for model selection occurred for models with process errors on recruitment and survival. Marginal AIC accuracy was poor for models with process errors on recruitment and natural mortality. Estimating the mean natural mortality rate had small effects on the accuracy of AIC in selecting the appropriate process error. Estimating the mean log-natural mortality resulted in a small decrease in marginal AIC accuracy for including the environmental effect. Marginal AIC was conservative for determining whether the environmental covariate affected natural mortality. Marginal AIC was very accurate in determining no effect when there was no effect in the operating model, but marginal AIC often ranked the null model best when there was an effect. Accuracy of marginal AIC for covariate effects improved with increased effect size, increased temporal contrast in the covariate, and lower uncertainty in observations.

Miller et al. (WP 5) found no evidence of bias in estimation of environmental effects regardless of process error assumptions when there was low uncertainty in the environmental observations and large contrast in the environmental covariate. In most cases the relative error of the estimated environmental effect did not depend on the source of process error assumed in the estimating model. The worst bias was observed when OMs assumed process errors on recruitment and survival, high uncertainty in covariate observations, low variability in the covariate, and low uncertainty in index and age composition observations. Simultaneously estimating the mean/intercept log natural mortality resulted in larger variation in the relative errors of the estimated environmental effect. Estimation of the intercept was reliable for all process error assumptions of the estimating model when the operating models assumed process errors on recruitment and natural mortality, contrast in fishing pressure over time, and lower observation error. Estimating the mean/intercept for log natural mortality generally resulted in highly variable estimates of annual natural mortality and spawning biomass and evidence of bias for some operating and estimation model assumptions about process error source. Again reliability of annual natural mortality estimates was generally improved with lower observation error uncertainty and contrast in fishing pressure.

Miller et al. (WP 5) found reliable detection of covariate effects requires informative data. Marginal AIC preferred simpler models than the true model when information content in data and contrast in covariates and abundance was low. The null model for environmental covariate effects (no covariate effect) was selected when contrast in the time series was low and/or uncertainty in observations was high. The selection of the null model by marginal AIC also likely decreases with strength of the effect of the covariate on M. Similarly, when there was process error in recruitment and natural mortality, estimation models with process error only in recruitment were preferred presumably due to low variation in simulated natural mortality process errors. Covariate effect estimation can be robust to process error assumptions with high contrast in covariate and low observation error.

3.7.4 Miller (WP 6)

State-space models are well suited to account for temporal variation in assessment model parameters that arises from environmental variation that cannot be attributable to explicit environmental covariates. This temporal variation results in equilibria that are no longer deterministic equations. Miller (WP 6) shows that the central tendency (mean/median) of stochastic equilibrium spawning stock biomass cannot be accurately estimated using analytical equilibrium methods. However, stochastic equilibrium age-specific contributions to spawning stock biomass for each age less than the plus group can be estimated accurately. The issue is that spawning stock biomass and the plus group are sums of log-normal random variables and there is no analytic result for the moments of this distribution. Using the bias-correction option in TMB does not generally solve this problem because of the size of the variance of the log-normal random variables being summed.

Miller (WP6) recommends distinguishing the analytic and stochastic reference points when reporting them. Which type of reference point performs better for management remains an open question. If accurate estimation of the age-specific components of the biomass reference point is sufficient, then it might be recommended to expand the population age structure to minimize the contribution of the plus group to the biomass or catch reference point.

3.8 Recommendations

1. Limit the investigations to a set of covariates that biology suggests close links of the covariate to the particular demographic parameter. General recommendation of avoiding data-mining and spurious correlation when fitting statistical models also applies to investigation of environmental effects on the dynamics of fish stocks within state-space assessment models.
2. Evaluate effects of covariates against models that have temporal variation in the parameter which the covariate is hypothesized to affect. This already typically occurs for effects on recruitment, but Miller et al. (2018) showed it also applies to other demographic parameters of fish stocks. This parallels the investigation of covariate effects in traditional regression models where reductions in residual variation or deviance are measured.
3. Assess the relative magnitude of observation error in environmental covariates. Low observation error is important in environmental covariate observation as well as in other assessment data (age composition and abundance indices). Miller et al. (WP 5) found that the ability to make inferences about effects of environmental covariates on natural mortality was weaker with higher observation error. Hart and Hansell (WP 2) found lower convergence rates for models with high observation error on environmental covariates when estimating time varying catchability.
4. Fix parameters describing environmental process variability where information is known. Britten et al. (WP 1) found high relative bias in estimating environmental process standard deviation when environmental correlation was high.
5. Avoid the ‘masking’ functional form when relating stock-recruitment relationships to an environmental covariate (until further work can diagnose issues). Britten et al. (WP 1) found a low rate of convergence for this model, low identifiability, and found high relative bias in stock recruitment parameter estimation.
6. Ensure good contrast in the environmental covariate(s). Miller et al (WP5) found that higher variation in the latent covariate improved the ability to make accurate inferences about environmental effects on natural mortality.

7. Conduct retrospective comparisons of models with and without covariate effects to confirm inferences are consistent as the number of years with observations changes. Miller et al. (2016) demonstrated this approach to help avoid the phenomenon of the perception of an effect or correlation of the covariate breaking down as new data are obtained (Myers 1998, Francis 2006).
8. Conduct self-tests as described in TOR 1 to confirm reliability of the estimation of effect size the covariate has on the assessment model parameter estimates and the reliability of biological reference points derived from them.

TOR 4. Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.

4.1 Introduction

Simulation models have been widely used in fisheries to test the impacts of different stock assessment and management approaches on achieving short- and long-term management objectives. Studies have compared different assessment model types (e.g., production vs. age-structured; Horbowy 2011), different types of model misspecification (e.g., Hordyk et al. 2019), different harvest control rules (e.g., Wildermuth et al. 2023), and the inclusion of environmental drivers in assessment models or management procedures (Punt et al. 2014, Haltuch et al. 2019). The advancement of state-space assessment models has the potential to not only improve the accuracy of assessment estimates, but also our ability to accurately identify environmental drivers of stock productivity, which could lead to improved management advice for a stock. There is general agreement that assessment models used in management should move toward using state-space approaches to estimation (Hoyle et al. 2020, Punt 2023), but it is also important to evaluate the effects of this added complexity on fisheries management (Peterson and Walter 2023). The effects of the state-space management model and whether it accounts for environmental effects would most naturally be evaluated through closed loop simulation studies that may also investigate a range of harvest control rules (Punt et al. 2016).

Many simulation studies have been conducted testing the utility of state-space assessment models compared to more traditional stock assessment approaches (e.g., Jiao et al 2012, Miller and Hyun 2018, Stock and Miller 2021, Stock et al. 2021). These studies have typically focused on comparison of assessment estimates across model types to the “true” values in the operating model. Few have done a full closed-loop simulation that compares the long-term performance of using complex state-space or other assessment approaches on achieving management objectives. Thus, there appears to have been little research conducted that would directly inform this Term of Reference given the relative novelty of state-space age-structured assessment models being applied in management.

Many management strategy evaluations (MSE) account for temporal variation in parameters in the operating model arising from explicit interactions of species (e.g., Punt and Butterworth 1995, Deroba et al. 2019), but it seems far fewer also attempt to account for this temporal variation in the estimating models because typically the estimating or management model is far less complex than the operating model (Kaplan et al. 2021). Similarly, effects of environmental covariates may be included in the operating model and/or estimating model (e.g., A'mar et al. 2009, Haltuch and Punt 2014).

Closed-loop simulation models have also been widely used to test the ability for assessment models to identify environmental drivers of stock productivity, and to incorporate such drivers into the management system through population projections or in harvest control rules. Punt et al. (2014) reviewed a number of these simulation studies, and found that, in general, simple management procedures tended to perform as good or better than more complex procedures that incorporate environmental drivers. Most studies in this area have focused on environmental drivers impacting recruitment, and Haltuch et al. (2019) reviewed studies that attempted to include environmental drivers in recruitment forecasts or management strategy evaluation simulations. They found that such attempts generally performed poorly, and only tended to work well for stocks with short pre-recruit windows, or distinct bottleneck periods where known drivers are more easily identified.

Although there have been recent applications of state-space estimation models (Hordyk et al. 2023, Su 2023), it appears none have used state-space age-structured estimation models to treat the temporal variation as random effects with marginal likelihood-based estimation. There is general agreement that assessment models used in management should move toward using state-space approaches to estimation (Hoyle et al. 2020, Punt 2023), but it is also important to evaluate the effects of this added complexity on fisheries management (Peterson and Walter 2023).

The WG did not have sufficient time to complete the closed-loop simulation studies using state-space and traditional assessment models as estimating models which would directly inform this Term of Reference. We describe the structure of such a simulation study in Section 4.2.

4.2 Simulations study structure to complete this Term of Reference

A suite of operating models should be configured analogous to those done for simulation studies described by Britten et al. (WP 1) and Miller et al. (WP 5) with a groundfish life history type, a fishing fleet, and spring and fall indices representing NEFSC bottom trawl surveys. The suite of state-space operating models should span a large number of years (e.g, 100), and have alternative assumptions about process errors on recruitment, survival, and natural mortality. The factors that should vary across operating models are

- the magnitude of observation errors (low to high),
- magnitude of variance and autocorrelation in the process errors (low to high), and
- fishing history (e.g., light, moderate, and heavy historical exploitation).
- alternative environmental covariate effects (none, small, large) or different random process errors (e.g., random walk, AR1, AR2) on recruitment, natural mortality, and/or catchability

Using a set of seeds unique to each operating model, simulate stochastic processes and observations over some historical period. Most assessments in the region have at least 40 years of data included in the fitting, so a reasonable historical period would be at least 40 years. The fishing history during the historical period could be stock-specific, or more generic, representing different patterns in the region (see Legault et al. 2023). Starting at year 40 of the operating model alternative WHAM estimation models to simulated observations up to year 40 would be fit. The alternative WHAM models should include

- alternative state-space model configurations (e.g., alternative process error assumptions),
- alternative assumptions about environmental effects on recruitment and natural mortality, and
- a statistical catch age age model without random effects, mimicking a traditional statistical catch at age model.

Assessments in the region are typically done every 2-3 years, with catch advice based on projections over the interval between assessments. Given the focus on understanding the impacts of the state-space model on management advice, the simulations should include a single harvest control rule to reduce the potential for confounding effects of different control rules. The current acceptable biological catch (ABC) control rule used in New England applies a target *F* of 75% of $F_{40\%}$ for most stocks. So, assuming three years between assessments:

- Conduct projections starting in year 41 through year 43 to determine the catch advice based on fishing at 75% of $F_{40\%}$ (alternative projection types may be considered given the flexibility in WHAM options),
- Set catch in the operating model for years 41 to 43 with corresponding annual *F* determined internally (i.e., no management uncertainty),
- Re-simulate the processes and observations given the assigned random seed.

Conduct an assessment in year 43, and repeat these steps for each assessment cycle up to the end of the time frame of the operating model (e.g., 100 years).

The simulation studies completed by the working group (e.g., WPs 1–5) found poor convergence of some estimation models and the same issue could arise in this closed-loop study. A step could be added to the management model to attempt a less complex state-space model in such simulations to mimic the likely real-world strategy.

Given the completed simulations, a range of performance metrics would be calculated that summarize the state of the fishery and the stock. Such metrics could include the average catch, interannual variation in catch, average stock biomass, proportion of time the stock is overfished (both based on the true stock size and perceived by the estimation model), and the proportion of years when overfishing occurs. Comparison of performance metrics would then be made across the different operating model factors for and among each alternative WHAM assessment model to quantify management performance of each assessment model, as well as the sensitivity and tradeoffs among metrics.

TOR 5: Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

5.1 Introduction

The NEFSC has had a strong production ageing program for many stocks for many years. This has allowed the use of age-based models for many of the primary stocks assessed in the region. The modeling approach for age-based models has evolved over time from virtual population analysis to statistical catch-at-age models, specifically the Age Structured Assessment Program ASAP, and now to state-space models (WHAM). The main reason for these changes in modeling approaches has been the development of computer software to allow easy implementation of more realistic and statistically sound approaches to stock assessment. An additional benefit of changing from ASAP to WHAM is that typically the retrospective patterns are substantially reduced, a problem confronting many assessments in the region. The ability to easily estimate random effects in stock assessments is made available by Template Model Builder in WHAM. WHAM also provides access to recently developed approaches such as one-step ahead residuals and a framework for statistical evaluation of whether inclusion of environmental covariates associated with a number of parameters is warranted or not.

The purpose of this Term of Reference is to allow some local stock assessments to switch from ASAP to WHAM without waiting for their next Research Track assessment. The current rules of Management Track assessments in the region do not allow changing the model used for a stock assessment, unless the stock is evaluated during a topic-based Research Track, such as this State-Space Research Track. The following stocks demonstrate both the ability to mimic the ASAP results as well as some improvements due to the transition of the model from ASAP to WHAM. The purpose here is not to provide an evaluation of the stock assessment, but rather an evaluation of the transition of the model from ASAP to WHAM, so that future Management Track assessments can use WHAM to provide management advice. The results of the models shown here will not be used for management advice directly; they will be reviewed during their next Management Track assessment.

5.2 Georges Bank winter flounder (Hansel WP 8)

The Georges Bank (GB) winter flounder stock is currently assessed with a Virtual Population Assessment (VPA). A WHAM model was developed (WP8) following research recommendations from the last two management track assessments (2020, 2022). The goal of WHAM was to help account for process error, poor cohort tracking and large retrospective patterns in the VPA.

The most recent assessment was completed in 2022 and uses data from 1982 to 2021. The catch at age input to the VPA consisted of combined U.S. and Canadian landings and discards from 1982-2021 for ages 1-6 with a 7+ age group (WP8 Figure 2). The VPA was calibrated using abundance at age indices from the NEFSC spring survey (1982-2021, ages 1-7) and fall NEFSC trawl surveys (1981-2020, ages 0-6 lagged forward one year and age) and the Canadian spring bottom trawl surveys (1987-2021, ages 1-7; WP8 Figure 3). Stock size was estimated for ages 2-6 in the terminal year+1. The natural mortality rate was assumed as 0.3 per year. A three year moving average was used to estimate maturity using data from the spring NEFSC survey (1982-2021).

The same data inputs were used as the 2022 management track assessment. No changes were explored regarding natural mortality, stock recruit relationship or environmental covariates; these will be explored in the upcoming research track assessment (2026). Several assumptions were made when moving from VPA to WHAM: 1) logistic selectivity for the commercial fleet and all three indices; 2) input effective sample sizes were assumed to be 200 for all data sources and years. High input effective sample sizes were needed for the self-weighting age composition distributions. Model exploration was done in an order: 1) age distributions; 2) recruitment assumptions; 3) time varying selectivity; 4) full state-space models. A set of common diagnostics were used to evaluate the different model runs: convergence, residuals, Akaike's Information Criteria (AIC), retrospective patterns, prediction skills and estimation performance.

WHAM like runs successfully converged and produced similar trends as the VPA for SSB, F, age ones and numbers at age. However, WHAM estimated slightly higher SSB and F since the mid-1990s (WP8 Figure 4-5). Selectivity estimates were similar for the commercial fleet and the three indices (WP8 Figure 4). Runs exploring the different age-comps supported using a logistic normal age distribution that ignored missing values; the runs using this distribution had some of the best Mohn's rho values for SSB and F (WP8 Table 1). Additionally, this age composition produced decent residual patterns for all data inputs (WP8 Figure 7-10). Thus, logistic normal age distribution was used in subsequent model runs.

Incorporating random effects into recruitment did not substantially improve model diagnostics so they were not included (WP8 Table 2). Models with random effects on fleet selectivity converged; however, model runs had similar residual patterns in the commercial age-comps and failed to improve Mohn's rho values (WP8 Table 3).

Full state-space models converged and including 2dar1 random effects on numbers at age led to similar AIC and improved Mohn's rho values for SSB and F (WP 8 Table 4; Figure 10-12).

Thus, the selected model had similar inputs to the VPA; however, has logistics-normal distribution on all age composition likelihoods and random effects on numbers at age using a 2dar1 process.

The selected WHAM run fit well to the commercial catch and the NEFSC surveys; however, struggled to fit to the DFO survey (WP 8 Figure 13-16). The selected WHAM run had a decent fit to all age composition data (WP 8 Figure 17-20).

The selected WHAM run had better retrospective patterns than the VPA and no longer requires a retrospective adjustment (WP 8 Figure 21-22). MASE scores were above one for each survey and were especially high for the Canadian survey (WP 8 Table 5; Figure 23). The selected model passed a self-test with mean bias below ten percent for R, SSB and F (WP 8 Table 6; Figure 24).

Biological reference points were similar between the VPA and WHAM (WP 8 Table 7). Stock status was also similar (WP 8 Figure 25). The F proxy reference point is most likely lower for WHAM because a logistic selectivity was used in WHAM, where the VPA had a dome in F at age. Projections were also similar between the two methods (WP 8 Table 8). Overall, trends in SSB and F were similar between the VPA and selected WHAM run (WP 8 Figure 26).

Results from the VPA and WHAM produced similar trends (Figure 26). However, there are several advantages to moving to WHAM: 1) VPA is an outdated assessment platform that does not assume error in the catch; 2) the logistic normal age composition likelihoods are self-weighting; 3) the 2dar1 correlations on numbers at age helps to account for changes in survivorship and autocorrelation in recruitment; 5) the WHAM model has better diagnostics and removes major retrospective patterns observed in the VPA.

5.3 Acadian redfish (Linton WP 9)

The most recent benchmark assessment for Acadian redfish (*Sebastes fasciatus*) occurred in 2008 as part of the 3rd Groundfish Assessment Review Meeting (GARM III; NEFSC 2008), using an application of the Age-Structured Assessment Program (ASAP; Legault 2012). Updates to the 2008 benchmark assessment model occurred in 2012 (NEFSC 2012), 2015 (NEFSC 2015), 2017 (NEFSC 2017), and 2020 (NEFSC 2022). The most recent assessment update occurred in 2023, using data through 2022 (NEFSC *In Prep* b). Importantly, the data sources used in the assessment model were reweighted to improve the fit to the observed survey indices at the end of the assessment time series.

The current assessment models total fishery catch as a combination of commercial landings and discards (WP 9, Figure 1). There is a gap in the commercial age composition data after 1985, with new years of age data being added only recently, starting in 2017 (WP 9, Figure 2). The Acadian redfish assessment incorporates the spring and fall NMFS bottom trawl survey (BTS) indices of abundance (WP 9, Figure 3). The spring and fall BTS indices both show a decrease in abundance at the end of the assessment time series. There is a gap in the spring BTS age composition data between 1980 and 1984, and again after 1990, with new years of age data being added only recently, starting in 2017 (WP 9, Figure 4). The fall BTS age composition data has only one missing year of data in 2020, (WP 9, Figure 5). As with the fishery age composition data, strong year classes can be tracked through time in the spring and fall BTS data.

The current Acadian redfish assessment estimates the numbers-at-age (NAA) in the first year as deviations from an exponential decline, with an associated coefficient of variation (CV) of 0.01. Age-1 recruitment is estimated using a Beverton-Holt model. The CVs on the annual recruitment deviations are set equal to 0.1 for 1913-1964, when only total catch data are available, before increasing linearly to 0.8 in 1969. The annual recruitment deviation CVs then decrease linearly from 0.8 in 2020 to 0.52 in 2022. The CV on the steepness parameter is set equal to 0.2, and the CV on the stock recruitment scaler is set equal to 0.6. Fishery selectivity is estimated for a single time block, 1913-2022, using age-specific parameters, with the selectivity for ages 10 and older fixed at 1. Spring and fall BTS selectivities are estimated using age-specific parameters. Spring BTS selectivity for ages 8 and older are fixed at 1. Fall BTS selectivity for ages 4 and older are fixed at 1. While model diagnostics have generally been considered good, the model failed to fit the spring and fall BTS indices at the end of the assessment time series, raising concerns with the peer review panels for the 2020 and 2023 management track assessments (NEFSC 2022, NEFSC *In Prep* b).

Six Woods Hole Assessment Model (WHAM) configurations were explored in this research track (WP 9), with the short term goal of replicating the 2023 management track ASAP model results, and the long term goal of improving the model fit to the survey indices in a future management track or research track assessment. Due to the tight constraints on key model parameters (i.e., NAA in the first year and recruitment) in the 2023 ASAP model, the initial WHAM configuration was set up with similar tight constraints on model parameters (WP 9, Table 1). Those constraints were incrementally loosened in the successive five model configurations.

Only five of the six WHAM configurations met the convergence criteria (WP 9, Table 1). From the converged models, the research track working group selected Model 5 as the preferred WHAM configuration for Acadian redfish. Model 5 estimates equilibrium NAA in the first year, Beverton-Holt recruitment with annual deviations treated as i.i.d. random effects, and logistic selectivity functions for the fishery and the fall and spring BTS indices. All five of the converged WHAM configurations produced similar fits to the observed data (WP 9, Figures 6-11, 13-18, 20-25, 27-32, and 34-39), and produced similar estimates of F, SSB, and R (WP 9, Figures 12, 19, 26, 33, and 40).

Model 5 had the lowest AIC score of the five models, but also had the highest Mohn's rho values for F, SSB, and R (WP 9, Table 2). The higher Mohn's rho values were likely due to the fact that Model 5 is estimating more parameters than Models 1-4. In conjunction with the improved AIC score, reducing the number of constraints on the model parameters was seen by the working group as a positive attribute of Model 5, even if it resulted in an increased retrospective pattern. The prediction error for Model 5, quantified by the Mean Absolute Scaled Error (MASE), was higher than the naïve method of predicting the future to be equal to the terminal year observation, with MASE scores equal to 1.02 and 1.17 for the fall and spring BTS indices, respectively. The relatively high MASE scores are likely due to the model's inability to fit the survey indices at the end of the assessment time series. The mean biases from a simulation self-test of Model 5 were relatively low, with a mean percent error of -0.4 % for F, 8.6% for R, and 1.4% for SSB. The relatively low mean bias estimates are likely due to the constraints that remain on key model parameters.

Fits to total catch, fall BTS index, and spring BTS index were similar between Model 5 and the 2023 ASAP model (WP 9, Figures 43-45). Residual patterns for catch-at-age, fall BTS age compositions, and spring BTS age compositions were similar between Model 5 and the 2023 ASAP model (WP 9, Figures 46-48). AIC scores cannot be compared between Model 5 and the 2023 ASAP model, because Model 5 includes random effects and the 2023 ASAP model does not. Model 5 had higher Mohn's rho values for F, SSB, and R compared to the 2023 ASAP model (WP 9, Table 2), which had a minor retrospective pattern.

Estimates of F, SSB, and R were similar between Model 5 and the 2023 ASAP model (WP 9, Figure 49), though there were several notable differences. The recruitment estimates differ for 1913-1964, where Model 5 allows greater variability in recruitment than the 2023 ASAP model. Model 5 estimates higher recruitment in 2020 than the 2023 ASAP model. Model 5 estimates higher SSB at the start of the assessment time series than the 2023 ASAP model.

Comparing biological reference points (WP 9, Table 4), the F_{MSY}-proxy of $F_{50\%}$ is identical for Model 5 and the 2023 ASAP model (0.037). The SSB at the F_{MSY}-proxy is higher for Model 5 (200,260 mt) compared to the 2023 ASAP model (184,322 mt).

Comparing 4-year projections (WP 9, Table 5), Model 5 projected SSB is slightly lower than the 2023 ASAP model projected SSB in 2023 and 2024, and slightly higher than the 2023 ASAP model projected SSB in 2025 and 2026. Model 5 projected catches are slightly lower than the 2023 ASAP model projected catches in 2024, and slightly higher than the 2023 ASAP model projected catches in 2025 and 2026.

5.4 Atlantic mackerel (Curti and Hansel WP 10)

Atlantic Mackerel is currently assessed using the Age Structured Assessment Program (ASAP, Legault C.M 2012). A WHAM model was developed here to (WP 10) following research recommendations from the 2023 spring Management Track Assessment. The goal of the WHAM model was to improve diagnostics and retrospective patterns.

The most recent ASAP model incorporated ages 1-10+, fishery and survey data from 1968-2022, and three index time series (WP 10 Figures 1-3). Combined Canadian and U.S. fishery catch (commercial and recreational) were modeled as one fishing fleet with constant selectivity over time. Fishery selectivity was assumed to be flat-topped with age-specific selectivity parameters fixed at one for ages 6+. The primary survey dataset is an index of spawning stock biomass that is developed from a dedicated egg survey in Canadian waters and the May/June MARMAP (1977-1988) and EcoMon (1992-2022) ecosystem surveys in U.S. waters. Additionally, the NEFSC spring bottom trawl survey is incorporated into the model and is treated as two distinct time series for the Albatross (1968-2008) and Bigelow (2009-2022) years. For these time series, only ages 3+ are used to develop the index because recent otolith microchemistry work demonstrated that these ages were most representative of the unit stock and not just the local spawning contingent (Redding et al. 2020). Trawl survey selectivity is fixed at one for age-3 with ages 4+ estimated parameters. Natural mortality was assumed to be 0.2

and constant over both time and age. Annual maturity ogives developed from Canadian samples representing the northern spawning contingent were used. This assumption is consistent with trends in the egg index, which indicate that the majority of the spawning stock is composed of individuals from the northern contingent.

In this study, the same data inputs were used as the 2023 management track assessment. No changes were explored regarding natural mortality or environmental covariates. Model exploration was completed in the following order: 1) age distributions; 2) recruitment assumptions; 3) time-varying selectivity; 4) full state-space models. A set of common diagnostics were used to evaluate the different model runs: convergence, residuals, Akaike's Information Criteria (AIC), retrospective patterns, prediction skills and estimation performance.

All ASAP material from the last Management Track assessment as well as current WHAM runs are available on GitHub: https://github.com/kcurti/Mack.WHAM.Development. See "run_notes.doc" on GitHub for a detailed description of each WHAM run.

A WHAM run configured the same way as ASAP successfully converged and produced similar trends as ASAP (WP 10; Figure 4 Run 8 on GitHub). Runs exploring the different age-compositions supported using a logistic normal age distribution that used an ar1 process to estimate zero values; the runs using this distribution resulted in similar temporal trends but had improved residual patterns and similar Mohn's rho values to other age composition likelihoods (WP Table 1; Figure 7). Thus, the logistic normal age distribution with an ar1 process to estimate zero values was used in subsequent model runs.

Incorporating random effects into recruitment alone did not substantially improve model diagnostics so they were not included (WP Table 2; Figure 8; Runs 4-6 on Github). Models fit with a Beverton-holt stock recruit relationship had similar diagnostics as models that assumed mean recruitment; however, the Beverton-holt stock recruit relationship did not fit well to observed larger SSB values (Run 6 on GitHub). Models with random effects on fleet selectivity converged but did not significantly improve diagnostics (WP 10 Table 3; Figure 9-10; Runs 8-10 on Github).

Full state-space models converged and including 2dar1 random effects on numbers-at-age led to improved Mohn's rho values (WP 10 Table 4; Figure 11-12; Run 12-15 on Github). Thus, the selected model had similar inputs to the original ASAP model; however, it had logistic-normal distributions with an ar1 process on all age composition likelihoods and random effects using a 2dar1 process on numbers-at-age (Run 13).

The selected WHAM run had similar diagnostics as ASAP (WP Figure 16-17). The proposed WHAM run had decent fits to the age composition data (WP Figure 18-20) and had better retrospective patterns than ASAP (WP Table 5; Figure 21). The MASE scores were comparable between the selected WHAM run and other runs (WP Table 2-4; Figure 22). The self-test on the selected model produced mean bias below 10% for F, SSB and R (WP Table 6; Figure 23). Reference points were similar between ASAP and WHAM (WP Table 7). Reference points from WHAM were similar to those from ASAP but smaller in magnitude (WP Table 7); however, stock status was robust to these differences (WP Figure 24). The proposed WHAM run has similar trends as ASAP; however, it estimates larger uncertainty in SSB and F trends (WP Figure 25). This increased uncertainty is presumably due to the large process error estimates for recruitment and survival (2dar1 random effects on NAA); however, future work should further evaluate the trade-offs between model diagnostics and uncertainty estimates.

Short-term projections were completed at $F_{msy\ proxy}$. These WHAM projections exhibited the similar temporal trends as those from the ASAP model, but had much larger confidence intervals associated with the projected estimates (WP 10 Table 8, Figure 26).

Overall, the proposed WHAM run produced similar trends as the previous ASAP model. There are several advantages from moving to WHAM: 1) the logistic-normal age compositions are self weighting; 2) including 2dar1 random effects on numbers-at-age reduces retrospective patterns; 3) the 2dar1 random effects on numbers-at-age also helps to account for recruitment correlations. However, the proposed WHAM run also has increased uncertainty around key parameter estimates. More work is needed to further determine if these uncertainties are realistic or an artifact of model specification. The base WHAM model that was configured like ASAP (fixed effects only) had similar trends to ASAP and similar estimates of uncertainty. At a minimum, moving over to WHAM and configuring it the same as ASAP would allow for future model development.

5.5 Gulf of Maine haddock (Perretti WP 11)

The Gulf of Maine (GOM) haddock stock is currently assessed using the Age Structured Assessment Program (ASAP, Legault C.M 2012). A WHAM was developed here (WP 11) following a research recommendation from the 2022 Research Track assessment (NEFSC *In Prep* a). The main motivation for developing a WHAM was to address a recent increase in retrospective error, and to improve model fit to the survey indices in the most recent decade.

The most recent assessment update occurred in 2022 using data through 2021 (NEFSC 2022). The assessment incorporates both the spring and fall NMFS Bottom Trawl Survey (BTS) (WP 11, Figures 1 & 2). There is a commercial and recreational fishery, which the model combines into a single fleet, and the combined age-composition shows similar strong year-classes as the surveys (WP 11, Figure 4).

Although the model diagnostics have been considered good overall in past reviews, the model has struggled to fit the rapid rise and fall in the observed survey indices in the last decade (WP 11, Figure 6). In particular, the high abundance years immediately following the 2013 year class, as well as the most recent decline in abundance as that year class has aged out. Relatedly, the model developed a major retrospective pattern in the 2019 update due to the rapid increase in SSB implied by the survey observations. In the most recent update, a moderate retrospective pattern has developed in the opposite direction as SSB has declined rapidly (NEFSC 2022).

An initial ASAP-like WHAM was explored to bridge from ASAP to WHAM. The ASAP-like WHAM run was configured to be as similar as possible to the accepted GOM haddock ASAP model. After that, three additional WHAM runs were tested that included 2d-ar1 random effects in numbers-at-age (NAA RE). The hypothesis was that including NAA random effects would give the model more flexibility, which may improve the fit to the survey and decrease the retrospective error. Each NAA RE run differed in the age-composition likelihood: (1) multinomial, (2) Dirichlet, (3) logistic-normal. In addition to working paper X, full diagnostic plots and *R* code used to perform the WHAM runs can be found at https://github.com/NEFSC/READ-PDB-PERRETTI_HADSSRT.

ASAP-like model

Convergence and an invertible hessian were achieved once population abundance initial year values for ages seven and eight were fixed to their starting values. Although the fit to the aggregate fishery catch was similar to the ASAP model, the fit to the survey time series was substantially worse in the terminal year of the WHAM ASAP-like model (WP 11, Figure 10). The ASAP model fits the terminal observation within the 90% confidence interval, however the WHAM ASAP-like model has a large residual in the terminal year. This residual is the largest in the time series for the spring survey, and the second largest in the time series for the fall survey.

In addition to the large terminal residuals in the survey indices, the ASAP-like model had a worse retrospective error in recruitment compared to the ASAP model (WP 11, Table 1). The increase in retrospective error was driven by a large overestimate of the 2019 year class strength when the terminal year is 2020 (WP 11, Figure 11). This also resulted in poor prediction accuracy in the terminal year, as shown in WP 11, Figure 12.

Numbers-at-age random effects models

As mentioned previously, due to the poor performance of the ASAP-like WHAM, three numbers-at-age random effect models were tested, each of which differed only in the age-composition likelihood used. They are labeled RE multinomial, RE Dirichlet, and RE logistic-normal.

The most notable difference was between the ASAP-like model and all of the RE models, where the RE models better fit the terminal index observation in both surveys (WP 11, Figure 13). All models closely fit the aggregate fleet catch time series (WP 11, Figure 14). Age-composition residuals did not show problematic patterning in either the in-sample or one-step-ahead residuals (WP 11, Figures 15 - 20).

Retrospective error for SSB and *F* was lower in all WHAM models than the ASAP model (WP 11, Table 1). Retrospective error for SSB was lowest in the RE multinomial model, and for *F* it was lowest in the RE Dirichlet model. None of the WHAM models would require a retrospective adjustment if this were a Management Track assessment (i.e., the retro-adjusted SSB and *F* values did not fall outside of their 90% confidence intervals).

Overall, prediction error, quantified by Mean Absolute Scaled Error, was similar across models, and all models had lower error for the spring survey than the fall survey. Although MASE scores were similar, plots of the predictions from each model show clear differences in the patterns of error (WP 11, Figures 12 & 21 - 23). As mentioned previously, the ASAP-like model had large errors in the terminal year due to an overestimate of the 2019 year class recruitment in 2020 (WP 11, Figure 12). The RE logistic-normal model suffered a similar overestimate of recruitment and subsequent over-prediction of survey abundance in the terminal year. In contrast, the RE multinomial (WP 11, Figure 22) and RE Dirichlet (WP 11, Figure 23) both correctly predict the decline in abundance in the terminal five years.

In simulation self-tests, all models had high convergence rates, however estimation error varied across models (WP 11, Table 3). The ASAP-like model had the lowest average error in all measured variables at 1%. All of the RE models had substantially higher error than the

ASAP-like model, with the highest error found in the RE logistic-normal model (26%) and the lowest error in the RE Dirichlet model (13%). The pattern of bias was similar in the RE models with all underestimating SSB in the last decade and overestimating *F* in the 1980s (WP 11, Figures 24 - 27). Of the RE runs, only the Dirichlet model captured the true values within the 90% confidence interval of the mean fit for all variables

All WHAM runs with random effects appear to be an improvement over the ASAP-like WHAM run, particularly due to their improved fit to the survey indices. Of the RE models, I recommend the RE Dirichlet model as the final model to replace the current ASAP model.

There are four main reasons that the RE Dirichlet model is preferable to the current ASAP model: First, the RE Dirichlet model has an improved fit to the survey time series, particularly in the final decade (compare WP 11 Figure 13 vs. Figure 10). Second, the RE Dirichlet model has substantially lower retrospective error in SSB (0.30 vs. 0.16) and *F* (-0.25 vs. -0.06) compared to the current ASAP model. Retrospective error on recruitment is higher in the RE Dirichlet model, however recruitment retrospective error is typically highly variable, and in this case is heavily influenced by a single outlier estimate in 2020. Third, the RE Dirichlet model was one of only two WHAM models that correctly predicted the direction of change in the survey indices in the final decade. Fourth, the RE Dirichlet model had the lowest self-simulation error of all random effect models (WP 11, Table 3).

Trends in SSB, F, and recruitment are similar between the RE Dirichlet model and the accepted ASAP model, with generally lower *F* and higher SSB in the RE Dirichlet model, and increased variability in the RE Dirichlet model (WP 11, Figure 31). The overfishing reference point is similar between the two models with RE Dirichlet having a slightly lower $F_{40\%SPR}$. $SSB_{40\%SPR}$ is higher in the RE Dirichlet model due to the higher scale of recruitment throughout the timeseries. Projections are similar overall between the two models.

References

Aeberhard, W.H., Mills Flemming, J. and Nielsen, A., 2018. Review of state-space models for fisheries science. Annual Review of Statistics and Its Application, 5, pp.215-235.

Albertsen, C. M., Nielsen, A., and Thygesen, U. H. 2017. Choosing the observational likelihood in state-space stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences. 74(5): 779-789. https://doi.org/10.1139/cjfas-2015-0532

A'mar, Z. T., Punt, A. E., & Dorn, M. W. (2009). The evaluation of two management strategies for the Gulf of Alaska walleye pollock fishery under climate change. ICES Journal of Marine Science, 66(7), 1614-1632.

Aregay, M., Shkedy, Z. and Molenberghs, G., 2013. A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. Computational Statistics & Data Analysis, 57(1), pp.233-245.

Auger-Méthé, M., Field, C., Albertsen, C.M., Derocher, A.E., Lewis, M.A., Jonsen, I.D. and Mills Flemming, J., 2016. State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Scientific reports, 6(1), p.26677.

Auger‐Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A.A., Leos‐Barajas, V., Mills Flemming, J., Nielsen, A., Petris, G. and Thomas, L., 2021. A guide to state–space modeling of ecological time series. Ecological Monographs, 91(4), p.e01470. doi: 10.1002/ecm.1470

Bates, N. R., & Peters, A. J. (2007). The contribution of atmospheric acid deposition to ocean acidification in the subtropical North Atlantic Ocean. Marine Chemistry, 107(4), 547-
558.

Berg, C.W. and Nielsen, A., 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES Journal of Marine Science, 73(7), pp.1788-1797.

Breivik, O. N., Aldrin, M., Fuglebakk, E., Nielsen, A. 2023. Detecting significant retrospective patterns in state space fish stock assessment. Canadian Journal of Fisheries and Aquatic Sciences 80(9): 1509-1518. doi:10.1139/cjfas-2022-0250

Britten, G.L., Dowd, M., Worm, B. 2015. Changing recruitment capacity in global fish stocks. Proceedings of the National Academy of Sciences 113 (1): 134-139

Brooks, E.N., and Legault, C.M. 2016. Retrospective forecasting – evaluating performance of stock projections for New England groundfish stocks. Can. J. Fish. Aquat. Sci. 73(6):935–950. doi:10.1139/cjfas-2015-0163.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., and Bolker, B. M. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R journal, 9(2), 378-400

Cadigan, N.G., 2016. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Canadian Journal of Fisheries and Aquatic Sciences, 73(2), pp.296-308.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R. and Maunder, M.N., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research, 240, p.105959.

Celeux G., Durand, J.-B. 2008. Selecting hidden Markov model state number with cross-Validated Likelihood. Computational Statistics 23 (4): 541-564. doi: 10.1007/s00180-007-0097-1

Conn, P.B., Williams, E.H. and Shertzer, K.W., 2010. When can we reliably estimate the productivity of fish stocks? Canadian Journal of Fisheries and Aquatic Sciences, 67(3): 511-523.

Conn, P. B. and Johnson, D. S. and Williams, P. J. and Melin, S. R. and Hooten, M. B. 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs 88(4): 526–542

Correa, G. M., Monnahan, C. C., Sullivan, J. Y., Thorson, J. T., & Punt, A. E. (2023). Modelling time-varying growth in state-space stock assessments. ICES Journal of Marine Science, 80(7), 2036-2049.

Crone, P. R., Maunder, M. N., Lee, H., & Piner, K. R. (2019). Good practices for including environmental data to inform spawner-recruit dynamics in integrated stock assessments: small pelagic species case study. Fisheries Research, 217, 122-132.

Cushing, D.H., 1990. Plankton production and year-class strength in fish populations: an update of the match/mismatch hypothesis. Advances in marine biology 26: 249-293.

De Oliveira, J. A. A., & Butterworth, D. S. (2005). Limits to the use of environmental indices to reduce risk and/or increase yield in the South African anchovy fishery. African Journal of Marine Science, 27(1), 191-203.

de Valpine, P. and Hastings, A., 2002. Fitting population models incorporating process noise and observation error. Ecological Monographs, 72(1), pp.57-76.

Dennis, B., Ponciano, J.M., Lele, S.R., Taper, M.L. and Staples, D.F., 2006. Estimating density dependence, process noise, and observation error. Ecological Monographs, 76(3), pp.323-341.

Deroba, J.J., Schueller, A.M., 2013. Performance of stock assessments with misspecified age- and time-varying natural mortality. Fish. Res. 146, 27–40

Deroba, J.J., Gaichas, S.K., Lee, M.Y., Feeney, R.G., Boelke, D. and Irwin, B.J., 2019. The dream and the reality: meeting decision-making time frames while incorporating ecosystem and economic models into management strategy evaluation. Canadian Journal of Fisheries and Aquatic Sciences, 76(7), pp.1112-1133.

Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence‐only data. Global Ecology and Biogeography, 23(12), pp.1472-1484.

Dorner, B., Peterman, R.M. and Haeseker, S.L., 2008. Historical trends in productivity of 120 Pacific pink, chum, and sockeye salmon stocks reconstructed by using a Kalman filter. Canadian Journal of Fisheries and Aquatic Sciences, 65(9): 1842-1866.

du Pontavice, H., Miller, T. J., Stock, B. C., Chen, Z., and Saba, V. S. 2022. Ocean model-based covariates improve a marine fish stock assessment when observations are limited. ICES Journal of Marine Science, 79(4), 1259-1273.

Durbin J., Koopman S.J. 2001. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.

Fasiolo, M., Pya, N. and Wood, S.N., 2016. A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. Statistical Science, pp.96-118.

Fay, G., & Punt, A. E. 2013. Methods for estimating spatial trends in Steller sea lion pup production using the Kalman filter. Ecological Applications *23*(6): 1455-1474.

Fong, Y., Rue, H., & Wakefield, J. 2010. Bayesian inference for generalized linear mixed models. Biostatistics, 11(3), 397-412

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2011. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods & Software 27, 233–249. doi: 10.1080/10556788.2011.597854

Fisch, N., Camp, E., Shertzer, K., & Ahrens, R. 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fisheries Research, 243, 106069.

Fisch, N., Shertzer, K., Camp, E., Maunder, M., Ahrens, R. 2023. Process and sampling variance within fisheries stock assessment models: estimability, likelihood choice, and the consequences of incorrect specification. ICES Journal of Marine Science. doi: 10.1093/icesjms/fsad138

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A. and Sibert, J., 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software, 27(2), pp.233-249.

Francis, R.C., 2006. Measuring the strength of environment–recruitment relationships: the importance of including predictor screening within cross-validations. ICES Journal of Marine Science 63(4): 594-599.

Francis, R.C., 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences, 68(6), pp.1124-1138.

Francis, R.C., 2014. Replacing the multinomial in stock assessment models: A first step. Fisheries Research, 151, pp.70-84.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. Chapman and Hall/CRC, London,UK

Gudmundsson, G. 1994. Time series analysis of catch-at-age observations. Applied Statistics 43: 117-126.

Haltuch, M. A., Punt, A. E. 2011. The promises and pitfalls of including decadal-scale climate forcing of recruitment in groundfish stock assessment. Canadian Journal of Fisheries and Aquatic Sciences, 68(5): 912–926. doi: 10.1139/f2011-030

Hansell, A., Walter, J., Cadrin, S., Golet, W., Hanke, A., Lauretta, M., & Kerr, L. 2020. Incorporating the Atlantic multidecadal oscillation into the western Atlantic bluefin tuna stock assessment. Collect. Vol. Sci. Pap. ICCAT, 77(2), 376-388.

Hare, J.A., Morrison, W.E., Nelson, M.W., Stachura, M.M., Teeters, E.J., Griffis, R.B., Alexander, M.A., Scott, J.D., Alade, L., Bell, R.J. and Chute, A.S., 2016. A vulnerability assessment of fish and invertebrates to climate change on the Northeast US Continental Shelf. PloS one, 11(2), p.e0146756.

Hare, J.A., 2014. The future of fisheries oceanography lies in the pursuit of multiple hypotheses. ICES Journal of Marine Science, 71(8), pp.2343-2356.

He, X., Field, J.C., Pearson, D.E. and Lefebvre, L.S., 2016. Age sample sizes and their effects on growth estimation and stock assessment outputs: Three case studies from US West Coast fisheries. Fisheries Research 180: 92-102.

Hjort , J. 1914. Fluctuations in the great fisheries of northern Europe viewed in the light of biological research. Journal du Conseil Permanent International pour L'Exploration de la Mer, 20: 1-228.

Hobbs et al. 2015. State-space modeling to support management of *brucellosis* in the Yellowstone bison population. Ecological Monographs 85(4): 522–556.

Hollowed, A. B., Ianelli, J. N., & Livingston, P. A. 2000. Including predation mortality in stock assessments: a case study for Gulf of Alaska walleye pollock. ICES Journal of Marine Science, 57(2): 279-293.

Hooten, M. B., Hobbs, N.T. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85(1): 3-28. doi: 10.1890/14-0661.1

Hordyk, A., Brown, C., Carruthers, T., Coelho, R., Duprey, N., Gillespie, K., Hanke, A., Miller, S., Rueda, L., Rosa, D. and Schirripa, M. 2023. Evaluation of performance of candidate management procedures for the North Atlantic swordfish management strategy evaluation. Collect. Vol. Sci. Pap. ICCAT, 80(1), pp.207-226.

Hoyle, S.D., Maunder, M.N., A'mar, Z.T., 2020. Frameworks for the next generation of general stock assessment models: 2019 CAPAM workshop report. N. Z. Fish. Assess. Rep. 2020/3 9. (http://www.capamresearch.org/sites/default/files/FAR-2020-39-CAPAM-Next-Generation-Stock-Assessment_Model-4115.pdf)

Hoyle, S.D., Maunder, M.N., Punt, A.E., Mace, P.M., Devine, J.A. and A'mar, Z.T., 2022. Preface: Developing the next generation of stock assessment software. Fisheries Research, 246: 106176. doi: 10.1016/j.fishres.2021.106176

Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. International journal of forecasting, 22(4), pp.679-688.

Ianelli, J. N., Hollowed, A. B., Haynie, A. C., Mueter, F. J., and Bond,N. A. 2011. Evaluating management strategies for eastern BeringSea walleye pollock (*Theragra chalcogramma*) in a changing environment.ICES Journal of Marine Science, 68: 1297–1304.

ICES. 2020. Workshop on the Review and Future of State Space Stock Assessment Models in ICES (WKRFSAM). ICES Scientific Reports. 2:32. 23 pp. http://doi.org/10.17895/ices.pub.6004

ICES. 2021. Workshop of Fisheries Management Reference Points in a Changing Environment (WKRPChange, outputs from 2020 meeting). ICES Scientific Reports. 3:6. 39 pp. https://doi.org/10.17895/ices.pub.7660

ICES. 2022. Workshop on ICES reference points (WKREF1). ICES Scientific Reports. 4:2. 70 pp. http://doi.org/10.17895/ices.pub.9749

Iles, T. C., & Beverton, R. J. H. (1998). Stock, recruitment and moderating processes in flatfish. Journal of Sea Research, 39(1-2), 41-55.

Jiao, Y., Smith, E. P., O'Reilly, R., and Orth, D. J. 2012. Modelling non-stationary natural mortality in catch-at-age models. ICES Journal of Marine Science, 69: 105–118.

Johnson, K.F., Monnahan, C.C., McGilliard, C.R., Vert-Pre, K.A., Anderson, S.C., Cunningham, C.J., Hurtado-Ferro, F., Licandeo, R.R., Muradian, M.L., Ono, K. and Szuwalski, C.S., 2015. Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. ICES Journal of Marine Science, 72(1), pp.137-150.

Kalman R.E. 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82: 35-45. doi:10.1115/1.3662552.

Kaplan, I.C., Gaichas, S.K., Stawitz, C.C., Lynch, P.D., Marshall, K.N., Deroba, J.J., Masi, M., Brodziak, J.K., Aydin, K.Y., Holsman, K. and Townsend, H., 2021. Management strategy evaluation: allowing the light on the hill to illuminate more than one species. Frontiers in Marine Science, 8, p.624355.

Kerr, L., Barajas, M., and Wiedenmann, J. 2022. Coherence and potential drivers of retrospective patterns in Northeast U.S. groundfish stock assessments. ICES Journal of Marine Science. doi:10.1093/icesjms/fsac140.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. 2016. TMB: automatic differentiation and Laplace approximation. Journal of Statistical Software 70(5), 1–21. doi: 10.18637/jss.v070.i05

Kuriyama, P.T., Ono, K., Hurtado-Ferro, F., Hicks, A.C., Taylor, I.G., Licandeo, R.R., Johnson, K.F., Anderson, S.C., Monnahan, C.C., Rudd, M.B. and Stawitz, C.C., 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, pp.119-127.

Lasker, R. 1981. The role of a stable ocean in larval fish survival and subsequent recruitment. pp. 80-87 *In:* Marine Fish Larvae: Morphology, Ecology and Relation to Fisheries (R. Lasker (Ed.)), Univ. Washington Press, Seattle

Lee, H.H., Maunder, M.N., Piner, K.R. and Methot, R.D., 2012. Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models? Fisheries Research, 125, pp.254-261.

Lee, Q., Thorson, J. T., Gertseva, V. V., & Punt, A. E. (2018). The benefits and risks of incorporating climate-driven growth variation into stock assessment models, with application to Splitnose Rockfish (*Sebastes diploproa*). ICES Journal of Marine Science, 75(1), 245-256.

Legault CM. 2012. Technical Documentation for ASAP Version 3.0 NOAA Fisheries Toolbox (https://noaa-fisheries-integrated-toolbox.github.io/).

Legault, C. M. and Restrepo, V. R. 1999. A flexible forward age-structured assessment program. Col. Vol. Sci. Pap. ICCAT 49(2): 246–253,

Legault, C. M., Wiedenmann, J, Deroba, J. J., Fay, G., Miller, T. J., Brooks, E. N., Bell, R. B., Langan, J. A., Cournane, J. M., Jones, A. W., Muffley, B. 2023. Data-rich but model-resistant: an evaluation of data-limited methods to manage fisheries with failed age-based stock assessments. Canadian Journal of Fisheries and Aquatic Sciences. 80(1): 27-42. doi: 0.1139/cjfas-2022-0045

Li, C., Deroba, J. J., Miller, T. J., Legault, C. M., Perretti, C. T. In review. An evaluation of common stock assessment diagnostic tools for choosing among state-space models with multiple random effects processes. Fisheries Research.

Liljestrand, E.M., Bence, J.R. and Deroba, J.J. 2023. Applying a novel state-space stock assessment framework using a fisheries-dependent index of fishing mortality. Fisheries Research 264: 106707.

Linton, B. C., and Bence, J. R. 2011. Catch-at-age assessment in the face of time-varying selectivity. ICES Journal of Marine Science, 68: 611-625.

Ludwig, D. and Walters, C.J. 1981. Measurement errors and uncertainty in parameter estimates for stock and recruitment. Canadian Journal of Fisheries and Aquatic Sciences 38(6): 711-720.

Magnusson, A. and Hilborn, R., 2007. What makes fisheries data informative? Fish and Fisheries, 8(4): 337-358.

MAFMC (Mid-Atlantic Fishery Management Council). 2019. Ecosystem Approach to Fisheries Management Guidance Document. https://www.mafmc.org/eafm.

Marandel, F., Lorance, P. and Trenkel, V.M., 2016. A Bayesian state-space model to estimate population biomass with catch and limited survey data: application to the thornback ray (*Raja clavata*) in the Bay of Biscay. Aquatic Living Resources, 29(2): 209

Marshall, K. N., Koehn, L. E., Levin, P. S., Essington, T. E., & Jensen, O. P. 2019. Inclusion of ecosystem information in US fish stock assessments suggests progress toward ecosystem-based fisheries management. ICES Journal of Marine Science 76(1): 1-9.

Martell, S. and Stewart, I., 2014. Towards defining good practices for modeling time-varying selectivity. Fisheries Research, 158, pp.84-95.

Maunder, M. N. and Deriso, R. B. 2003. Estimation of recruitment in catch-at-age models. Canadian Journal of Fisheries and Aquatic Sciences 60(10): 1204-1216. doi: 10.1139/f03-104

Maunder, M. N. and Punt, A. E. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries research, 70(2-3), 141-159.

Maunder, M. N. and Deriso, R. B. 2010. Dealing with missing covariate data in fishery stock assessment models. Fisheries Research 110(1-2): 80-86. doi: 10.1016/j.fishres.2009.09.009

Maunder, M.N. and Thorson, J.T. 2019. Modeling temporal variation in recruitment in fisheries stock assessment: a review of theory and practice. Fisheries Research, 217, pp.71-86.

Mazur, M. D., Jesse, J., Cadrin, S. X., Truesdell, S. B., & Kerr, L. 2023. Consequences of ignoring climate impacts on New England groundfish stock assessment and management. Fisheries Research, 262, 106652.

McAllister, M.K. and Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences, 54(2), pp.284-300.

Mendelssohn, R., 1988. Some problems in estimating population sizes from catch-at-age data. Fishery Bulletin, 86(4), pp.617-630.

Methot, R. D. and Wetzel, C. R. 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fisheries Research 142(1): 86–99.

Miller, T.J., Hare, J.A. and Alade, L.A., 2016. A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to southern New England yellowtail flounder. Canadian Journal of Fisheries and Aquatic Sciences 73(8), pp.1261-1270.

Miller, T.J. and Hyun, S.Y., 2018. Evaluating evidence for alternative natural mortality and process error assumptions using a state-space, age-structured assessment model. Canadian Journal of Fisheries and Aquatic Sciences, 75(5), pp.691-703.

Miller, T.J., O'Brien, L. and Fratantoni, P.S., 2018. Temporal and environmental variation in growth and maturity and effects on management reference points of Georges Bank Atlantic cod. Canadian Journal of Fisheries and Aquatic Sciences, 75(12), pp.2159-2171.

Minto, C., Mills Flemming, J., Britten, G. L., & Worm, B. 2014. Productivity dynamics of Atlantic cod. Canadian Journal of Fisheries and Aquatic Sciences, 71(2), 203-216.

Monnahan, C.C., Ono, K., Anderson, S.C., Rudd, M.B., Hicks, A.C., Hurtado-Ferro, F., Johnson, K.F., Kuriyama, P.T., Licandeo, R.R., Stawitz, C.C. and Taylor, I.G., 2016. The effect of length bin width on growth estimation in integrated age-structured stock assessments. Fisheries Research, 180: 103-112.

Monnahan CC, 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(5): e0197954. https://doi.org/10.1371/journal.pone.0197954

Myers, R.A., 1998. When do environment–recruitment correlations work? Reviews in Fish Biology and Fisheries 8: 285-305.

NEFSC (Northeast Fisheries Science Center). 2008. Assessment of 19 Northeast groundfish stocks through 2007. Report of the 3rd Groundfish Assessment Review Meeting (GARM III), Northeast Fisheries Science Center, Woods Hole, Massachusetts. August 4-8, 2005. NMFS NEFSC Ref. Doc. 08-15. 884 p.

NEFSC (Northeast Fisheries Science Center). 2012. Assessment or Data Updates of 13 Northeast Groundfish Stocks through 2010. US Department of Commerce, Northeast Fisheries Science Center Ref. Doc. 12-06; 789 p.

NEFSC (Northeast Fisheries Science Center). 2015. Operational Assessment of 20 Northeast Groundfish Stocks, Updated Through 2014. US Department of Commerce, Northeast Fisheries Science Center Ref. Doc. 15-24; 251 p.

NEFSC (Northeast Fisheries Science Center). 2017. Operational Assessment of 19 Northeast Groundfish Stocks, Updated Through 2016. US Department of Commerce, Northeast Fisheries Science Center Ref. Doc. 17-17; 259 p.

NEFSC (Northeast Fisheries Science Center). 2022. Management Track Assessments Fall 2022. NMFS NEFSC Tech. Memo. 305. https://doi.org/10.25923/380j-t283

NEFSC (Northeast Fisheries Science Center). *In Prep* a. Summary report of the Gulf of Maine haddock research track stock assessment. NMFS NEFSC Ref. Doc. XX-XX. XXX p.

NEFSC (Northeast Fisheries Science Center). *In Prep* b. Fall Management Track Assessments 2023. US Department of Commerce, Northeast Fisheries Science Center Ref. Doc. XX-XX; XXX p.

Newman, K.B., Buckland, S.T., Lindley, S.T., Thomas, L. and Fernandez, C., 2006. Hidden process models for animal population dynamics. Ecological applications, 16(1), pp.74-86.

Nielsen, A. and Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research, 158, pp.96-101.

Nye, J. A., Link, J. S., Hare, J. A., & Overholtz, W. J. 2009. Changing spatial distribution of fish stocks in relation to climate and population size on the Northeast United States continental shelf. Marine Ecology Progress Series, 393, 111-129.

Pepin, P., King, J., Holt, C., Gurney-Smith, H., Shackell, N., Hedges, K., & Bundy, A. 2022. Incorporating knowledge of changes in climatic, oceanographic and ecological conditions in Canadian stock assessments. Fish and Fisheries 23(6): 1332-1346.

Perreault, A.M. and Cadigan, N.G., 2021. Natural mortality diagnostics for state-space stock assessment models. Fisheries Research, 243, p.106062.

Perretti, C.T., Deroba, J.J. and Legault, C.M., 2020. Simulation testing methods for estimating misreported catch in a state-space stock assessment model. ICES Journal of Marine Science, 77(3), pp.911-920.

Peterman, R.M., Pyper, B.J. and Grout, J.A., 2000. Comparison of parameter estimation methods for detecting climate-induced changes in productivity of Pacific salmon (*Oncorhynchus* spp.). Canadian Journal of Fisheries and Aquatic Sciences, 57(1), pp.181-191.

Peterson CD. Walter III JF. 2023. Southeast Fisheries Science Center Management Strategy Evaluation Strategic Plan. NOAA Tech. Memo. NMFS-SEFSC-TM-766, 27 p. https://doi.org/10.25923/khnf-vh41.

Pinsky, M. L., Worm, B., Fogarty, M. J., Sarmiento, J. L., & Levin, S. A. (2013). Marine taxa track local climate velocities. Science, 341(6151), 1239-1242.

Punt, A. E. 2023. Those who fail to learn from history are condemned to repeat it: A perspective on current stock assessment good practices and the consequences of not following them. Fisheries Research 261: 106642. doi: 10.1016/j.fishres.2023.106642

Punt, A.E. and Butterworth, D.S., 1995. The effects of future consumption by the Cape fur seal on catches and catch rates of the Cape hakes. 4. Modelling the biological interaction between Cape fur seals Arctocephalus pusillus pusillus and the Cape hakes *Merluccius capensis* and *M. paradoxus*. South African Journal of Marine Science 16(1): 255-285.

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A. and Haddon, M., 2016. Management strategy evaluation: best practices. Fish and Fisheries 17(2): 303-334.

Punt, A. E., A'mar, T., Bond, N. A., Butterworth, D. S., de Moore, C. L., De Oliveira, J. A. A., Haltuch, M. A., *et al.* 2014. Fisheries management under climate and environmental uncertainty: control rules and performance simulation. ICES Journal of Marine Science, 71: 2208–2220.

Punt, A.E., Dunn, A., Elvarsson, B.Þ., Hampton, J., Hoyle, S.D., Maunder, M.N., Methot, R.D. and Nielsen, A., 2020. Essential features of the next-generation integrated fisheries stock assessment package: a perspective. Fisheries Research, 229, p.105617.

Pyper, B. J., Brian, R. M. Peterman. 1998. Comparison of methods to account for autocorrelation in correlation analyses of fish data. CJFAS, 55: 2127-2140

Ramos P., Oliveira, J.M. 2016. A Procedure for Identification of Appropriate State Space and ARIMA Models Based on Time-Series Cross-Validation. Algorithms 9(4):76. doi: 10.3390/a9040076

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F. and Dormann, C.F. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40: 913-929. https://doi.org/10.1111/ecog.02881

Romakkaniemi, A. 2015. Best practices for the provision of prior information for Bayesian stock assessment. ICES Cooperative Research Report No. 328, 93pp. https://doi.org/10.17895/ices.pub.5496

Saba, V.S., Griffies, S. M., Anderson, W. G., Winton, M., Alexander, M. A., Delworth, T. L., Hare, J. A., Harrison, M. J., Rosati, A., Vecchi, G. A., and Zhang, R. 2016. Enhanced warming of the Northwest Atlantic Ocean under climate change. Journal of Geophysical Research: Oceans, 121(1), 118-132. https://doi.org/10.1002/2015JC011346

Sampson, D.B., 2014. Fishery selection and its relevance to stock assessment and fishery management. Fisheries Research, 158, pp.5-14.

Schirripa, M. J., Goodyear, C. P., and Methot, R. M. 2009. Testing different methods of incorporating climate data into the assessment of US West Coast sablefish. ICES Journal of Marine Science, 66: 1605-1613.

Schirripa, M.J., Abascal, F., Andrushchenko, I., Diaz, G., Mejuto, J., Ortiz, M., Santos, M.N. and Walter, J., 2017. A hypothesis of a redistribution of North Atlantic swordfish based on changing ocean conditions. Deep Sea Research Part II: Topical Studies in Oceanography 140: 139-150.

Schnute, J.T. 1994. A general framework for developing sequential fisheries models. Can. J. Fish. Aquat. Sci. 51(8): 1676–1688. doi:10.1139/f94-168.

Sculley, M., Ijima, H., Chang, Y.-J., 2018. A base-case model in Stock Synthesis 3.30 for the 2018 north Pacific swordfish (*Xiphias gladius*) stock assessment. PIFSC working paper WP-18-005. http://doi.org/10.7289/V5/WP-PIFSC-18-005.

Skern-Mauritzen, M., Ottersen, G., Handegard, N.O., Huse, G., Dingsør, G.E., Stenseth, N.C. and Kjesbu, O.S. 2016. Ecosystem processes are rarely included in tactical fisheries management. Fish Fish, 17: 165-175. https://doi.org/10.1111/faf.12111

Silvar-Viladomiu, P., Minto, C., Brophy, D., & Reid, D. G. 2022. Peterman's productivity method for estimating dynamic reference points in changing ecosystems. ICES Journal of Marine Science 79(4): 1034-1047.

Solin, A. and Särkkä, S. 2015. State Space Methods for Efficient Inference in Student-t Process Regression. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. Eds: Lebanon, G. and Vishwanathan, S. V. N. Proceedings of Machine Learning Research 38: 885--893.

Stan Development Team. 2024. RStan: the R interface to Stan. R package version 2.32.5, https://mc-stan.org/.

Stawitz, C.C. and Essington, T.E., 2019. Somatic growth contributes to population variation in marine fishes. Journal of Animal Ecology, 88(2), pp.315-329.

Stock, B.C. and Miller, T.J., 2021. The Woods Hole Assessment Model (WHAM): a general state-space assessment framework that incorporates time-and age-varying processes via random effects and links to environmental covariates. Fisheries Research, 240: 105967.

Stock, B.C., Xu, H., Miller, T.J., Thorson, J.T. and Nye, J.A., 2021. Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for Southern New England-Mid Atlantic yellowtail flounder. Fisheries Research, 237, p.105873.

Stone, M. 1977. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. Journal of the Royal Statistical Society. Series B 39(1): 44-47.

Su, Z., 2023. Evaluation of management performance of a new state-space model for pink salmon (*Oncorhynchus gorbuscha*) stock-recruitment analysis. Canadian Journal of Fisheries and Aquatic Sciences 80:1268–1288. doi: 10.1139/cjfas-2022-0262

Subbey, S., Devine, J.A., Schaarschmidt, U. and Nash, R.D., 2014. Modelling and forecasting stock–recruitment: current and future perspectives. ICES Journal of Marine Science, 71(8): 2307-2322.

Sullivan, P.J., 1992. A Kalman filter approach to catch-at-length analysis. Biometrics, pp.237-257.

Szuwalski, C.S., Vert-Pre, K.A., Punt, A.E., Branch, T.A., Hilborn, R., 2015. Examining common assumptions about recruitment: a meta-analysis of recruitment dynamics for worldwide marine fisheries. Fish Fish. 16, 633–648.

Thorson, J.T., Shelton, A.O., Ward, E.J., Skaug, H.J., 2015. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. ICES J. Mar. Sci. J. Cons. 72(5), 1297–1310. doi:10.1093/icesjms/fsu243.

Thorson, J.T., Johnson, K.F., Methot, R.D. and Taylor, I.G., 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research, 192, pp.84-93.

Thorson, J. T. (2019). Measuring the impact of oceanographic indices on species distribution shifts: The spatially varying effect of cold‐pool extent in the eastern Bering Sea. Limnology and Oceanography, 64(6), 2632-2645.

Thorson, J.T., Miller, T.J. and Stock, B.C., 2023. The multivariate-Tweedie: a self-weighting likelihood for age and length composition data arising from hierarchical sampling designs. ICES Journal of Marine Science, 80(10), pp.2630-2641.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K. and Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics, 24, pp.317-339.

Trijoulet, V., Fay, G. and Miller, T.J., 2020. Performance of a state‐space multispecies model: What are the consequences of ignoring predation and process errors in stock assessments?. Journal of Applied Ecology, 57(1), pp.121-135.

Trijoulet, V., Albertsen, C.M., Kristensen, K., Legault, C.M., Miller, T.J. and Nielsen, A., 2023. Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fisheries Research, 257, p.106487.

Vaida, F. and Blanchard, S., 2005. Conditional Akaike Information for Mixed-Effects Models. Biometrika, pp.351-370.

Walters, C. J., & Hilborn, R. 1976. Adaptive control of fishing systems. Journal of the Fisheries Board of Canada, 33(1), 145-159.

Walters, C.J. and Ludwig, D., 1981. Effects of measurement errors on the assessment of stock–recruitment relationships. Canadian Journal of Fisheries and Aquatic Sciences, 38(6), pp.704-710.

Wiedenmann, J., and Jensen, O. P. 2018. Uncertainty in stock assessment estimates for new england groundfish and its impact on achieving target harvest rates. Canadian Journal of Fisheries and Aquatic Sciences,75: 342–356.

Wilberg, M. J., Thorson, J. T., Linton, B. C., & Berkson, J. 2009. Incorporating time-varying catchability into population dynamic stock assessment models. Reviews in Fisheries Science, 18(1): 7-24.

Wildermuth, R.P., Tommasi, D., Kuriyama, P., Smith, J. and Kaplan, I., 2023. Evaluating robustness of harvest control rules to climate-driven variability in Pacific sardine recruitment. Canadian Journal of Fisheries and Aquatic Sciences. doi: 10.1139/cjfas-2023-016

Xu, H., Miller, T.J., Hameed, S., Alade, L.A. and Nye, J.A., 2018. Evaluating the utility of the Gulf Stream Index for predicting recruitment of Southern New England-Mid Atlantic yellowtail flounder. Fisheries Oceanography, 27(1), pp.85-95.

Yin, J., Schlesinger, M. E., & Stouffer, R. J. (2009). Model projections of rapid sea-level rise on the northeast coast of the United States. Nature Geoscience,2(4), 262-266.

Zucchini, W., MacDonald, I. L., Langrock. R. 2016. Hidden Markov Models for Time Series: An Introduction Using R. CRC Press Boca Raton. 370pp.

Zhu, H., Vannucci, M. and Cox, D.D., 2010. A Bayesian hierarchical model for classification with selection of functional predictors. Biometrics, 66(2), pp.463-473.

Appendix 1: List of state-space research track working group members

Name	Institution	Contact
Andrew Applegate	New England Fisheries Management Council	aapplegate@nefmc.org
Greg Britten	MIT/Woods Hole Oceanographic Institution	gregory.britten@whoi.edu
Elizabeth N. Brooks	Northeast Fisheries Science Center	liz.brooks@noaa.gov
Gavin Fay	University of Massachusetts Dartmouth	gfay@umassd.edu
Alex Hansell	Northeast Fisheries Science Center	alex.hansell@noaa.gov
Christopher M. Legault	Northeast Fisheries Science Center	chris.legault@noaa.gov
Timothy J. Miller (Chair)	Northeast Fisheries Science Center	timothy.j.miller@noaa.gov
Brandon Muffley	Mid-Atlantic Fishery Management Council	bmuffley@mafmc.org
John Wiedenmann	Rutgers University	john.wiedenmann@gmail.com

Appendix 2: External participants and presenters

Name	Institution
Charles Adams	Northeast Fisheries Science Center
Gregory Bopp	Northeast Fisheries Science Center
Kristan Blackhart	NOAA Fisheries (HQ)
Russ Brown	Northeast Fisheries Science Center
Noel Cadigan*	Memorial University (Canada)
Giancarlo Correa	University of Washington/AZTI (Spain)
Kiersten Curti	Northeast Fisheries Science Center
Jon Deroba	Northeast Fisheries Science Center
Amanda Hart	Gulf of Maine Research Institute/Northeast Fisheries Science Center
Andrea Havron	NOAA Fisheries (HQ)
Lisa Kerr	Gulf of Maine Research Institute/University of Maine
Chengxue Li	Northeast Fisheries Science Center
Brian Linton	Northeast Fisheries Science Center
Emily Liljestrand	Michigan State University/Northeast Fisheries Science Center
Cole Monnahan	Alaska Fisheries Science Center
Brian Stock	Northeast Fisheries Science Center/Institute of Marine Research (Norway)
Angelia Miller	University of Massachusetts Dartmouth
Anders Nielsen*	Danish Technical University
Charles Perretti	Northeast Fisheries Science Center
Catalina Roman	University of Massachusetts Dartmouth
Michele Traver	Northeast Fisheries Science Center

* Presenter

Appendix 3: List of Contributed Working Papers with summaries

WP 1: Britten G, Brooks, E. N., Miller, T. Factors affecting estimation of environmental effects on recruitment in state space assessment models.

The authors completed a large scale simulation study with 512 operating models, each with 100 simulated data sets, and 6 estimated models fit to each simulated data set. The factors defining operating model configuration included the degree of variation in recruitment, temporal variation and autocorrelation of the environmental covariate, the uncertainty in indices and age composition observations, fishing history, and the magnitude of the effect of the covariate on recruitment, and the functional form relating the environmental covariate to recruitment. The estimating models make alternative assumptions on whether to include the environmental effect and which functional form to fit.

The authors found high convergence rates of the estimating models (>95%) except cases with low recruitment variability and constant fishing history. In >20% of these cases, the model failed to fit and caused R to abort. The authors look forward to follow-up work to better understand the cause.

In general there was low identifiability of an underlying stock recruitment model (33% correct); although rates of identification increased to above 50% with high contrast in fishing history and low recruitment variability. Identifiability was also low for the correct functional form of environmental covariate relationship (29%). Rates of correct identification depended on the strength of the environmental effect where higher effect sizes led to higher identification rates, and recruitment standard deviation where lower standard deviation led to high identification rates. Highest identification rates occurred for mean recruitment with no environmental covariate, exceeding 80%. Correct identification of both the stock-recruit relationship and environmental covariate relationship was very poor (10% overall), but was most successful with high contrast in fishing, low recruitment variability, and the null environmental covariate relationship.

All estimating models tended to fit the data similarly well, resulting in the average difference in AIC of 2.8 between the best and second-best fitting models (1.7 if recruitment variability was high), indicating support for many of the estimating models applied to the same data simulated from a given operating model.

Relative error was summarized over all years, the last 10 years, and the final year. Estimates were generally median unbiased and didn't differ among estimating models (the misspecified estimating models did as well as the correct estimating model). Higher recruitment variance led to wider range of relative error in model estimates of recruitment, but less so for spawning biomass and fishing mortality. In generalized linear (mixed) model-based effect size estimation, recruitment relative error depended most strongly on recruitment standard deviation and observation error where lower variance led to lower relative error, but the effects were generally small. The effect of recruitment random effect correlation increased for the final year where high correlation led to more positive relative bias. Spawning biomass followed the same general pattern as recruitment. Relative error for fishing mortality decreased with a high recruitment standard deviation.

Parameter bias showed interesting patterns, with recruitment standard deviation and fishing history playing the largest role, with exceptions. Parameters were median unbiased except for random effects process parameters. Latent environmental covariate standard deviation had positive relative bias when simulated correlation of latent covariates was high. Recruitment random effects parameters had bias that traded off - relative bias for recruitment standard deviation was positive when recruitment correlation was high, whereas recruitment correlation relative bias was negative when simulated recruitment correlation was low. Environmental effect size was the most poorly estimated parameter with the quantile range of relative error exceeding 20.

Median Mohn's rho was approximately zero, but the range of values increased when recruitment variance was high and there was low observation error. For spawning biomass and fishing mortality, there was a slight negative bias in median Mohn's rho for spawning biomass and slight positive bias in median Mohn's rho for fishing mortality. The retrospective pattern on the recruitment random effects generally had negative bias and a much wider range of values compared to recruitment, spawning biomass, and fishing mortality. The random effect retrospective was not lower for the correctly specified estimating model.

There was no discernable difference in the projections relative to two assumptions about the environmental covariate in the future (continue the estimated process or project at a value calculated as the mean of the last 5 years).

128 OMs of the 512 were simulated with a Beverton-Holt (BH) stock recruitment relationship (SRR) without an environmental effect. Two EMs were fit to those OMs that did not include an environmental effect - one with mean recruitment and one with a BH SRR. The authors found poor identifiability of the BH SRR based on AIC where mean recruitment was favored in the majority of cases despite an underlying BH SRR. The situation improved with high contrast in fishing history and low recruitment standard deviation. The authors also found that both SRR and mean recruitment EMs generally yielded unbiased assessment quantities (estimated R, SSB, and F); however, they did find minor effects of how OM factors contributed to relative error. Specifically, the effects of recruitment standard deviation and fishing history were anticorrelated among EMs with and without an SR relationship. High recruitment standard deviation and MSY fishing history generally led to more-positive relative bias for EMs without an SRR and led to more negative relative bias in EMs. This was more pronounced when assessing bias in the last year of the assessment.

WP 2: Hart, A.R. and Hansell, A. Factors affecting estimation of environmental effects on catchability in state-space assessment models.

Stock assessments typically assume that changes in catchability track changes in stock size, but this assumption may be violated when environmentally-driven shifts in distribution alter availability to the survey independent of stock size and temporal differences may not be well characterized by static catchability parameters. State-space models provide a path to account for time-varying catchability, but there are few examples where environmentally-driven changes in catchability have been incorporated directly into stock assessments.This simulation study evaluates state-space assessment performance across a range of observation uncertainties, fishing histories, and environmental effect sizes for models with the following catchability assumptions: 1) time-varying, 2) environmentally-driven, 3) both time-varying and environmentally-driven, and 4) constant (status quo) catchability. In total, 384 operating models were used for the different simulations. This study found that models with catchability random effects (both with and without environmental covariates) had lower convergence rates and less precise estimates for some model outputs compared to other assessment approaches, but their performance was robust against a wider range of model misspecifications. Status quo models had similar performance to other models and were preferentially selected by AIC when environmental effect size was zero and when catchability was misspecified for both seasons. However, status quo methods more regularly struggled to estimate a constant catchability that well characterized the underlying climate-driven trends, and these models were often more likely to generate biased estimates of key model results. Assessment models with only catchability covariates consistently had high convergence rates and were regularly selected by AIC when seasonal impacts were correctly specified. Overall, our results suggest that incorporating random effects or environmental covariates into stock assessment models improves their ability to characterize time-varying catchability when there is a moderate or strong environmental trend. When environmental relationships are well understood and seasonally correct, models with only an environmental covariate are able to account for time varying catchability. However, if the environmental trend is not well understood or is seasonally misspecified, random effects are more effective at accounting for time varying catchability.

WP 3: Li, C., Deroba, J. J., Miller, T. J., Legault, C. M., Perretti, C. T. An evaluation of common stock assessment diagnostic tools for choosing among state-space models with multiple random effects processes. (Also in review at Fisheries Research)

State-space models have received increasing attention in fisheries stock assessments given their flexibility to incorporate multiple sources of process errors. Identifying which process errors to include is important because incorrectly including some process errors can induce bias in management quantities. Existing model selection tools commonly applied in traditional statistical catch-at-age models may not perform as well for state-space models. We evaluated the efficacy of common diagnostic tools for correctly identifying the presence, absence, and magnitude of three process errors (survival, selectivity, and natural mortality) in a simulation–estimation experiment. No model diagnostic tools could consistently identify the correct process error structure in all situations. Incorrectly attributing the process error from natural mortality to other processes, or vice versa, led to relatively large bias in management quantities. Furthermore, incorrectly including an additional source of process error in the assessment models exhibited similar performance to the correct model and generally showed unbiased estimates of management quantities; incorrectly excluding a source of process error, however, generated large biases. Thus, despite not having generally reliable model diagnostic tools for state-space assessments, practitioners should err on the side of using overly complex models, except for natural mortality unless there is external corroborating evidence of changing M.

WP 4: Miller T. J., Applegate, A., Britten, G., Brooks, E. N., Fay, G., Hansell, A., Legault, C. M., Muffley, B., Wiedenmann, J. Factors affecting the reliability of state-space assessment models with alternative assumptions on sources of process errors.

We conducted a simulation study where we simulated process errors and observations for 72 operating models with alternative assumptions about fishing history, degree of uncertainty in index and age composition observations, type (recruitment, survival, fishery selectivity, catchability and natural mortality), degree of variation, and correlation of process errors. We fit 20 different estimating models to each of 100 simulated set of observations with alternative assumptions type and correlation structure of process errors, (mean) natural mortality was known or estimated, and a B-H stock recruit relationship was assumed or not.

Across simulations we summarized probability of convergence of fitted models, accuracy of marginal AIC in determining the correct process error assumption and it ability to determine the Beverton-Holt stock recruit relationship, bias in annual spawning stock biomass, in estimation of natural mortality, and in stock-recruit relationship parameters, and severity of retrospective patterns for estimation models.

Alternative measures of convergence performed differently. Invertible hessians and resulting standard error estimation was possible when criteria based on the gradient of the optimized log-likelihood with respect to the fixed effects parameters failed. Using hessian-based convergence, probability of convergence was best for models that assumed the correct source of process error, assumed M was known, and did not assume stock-recruit relationships.

Using marginal AIC provided most accurate inferences about the process errors on recruitment and survival and selectivity, in that AIC preferred EMs with assumed process errors that matched OMs, and EMs with those assumed process errors were not preferred when alternative process errors were assumed in the OMs. However, when the true process errors were more variable, AIC accuracy increased to a useful level. We found AIC more accurately determined a B-H stock recruit relationship rather than the null model without a S-R relationship when there was low variability in recruitment, low variability in survival random effects, and higher variation in spawning biomass over the time series. When the (mean) natural mortality rate was estimated, we found large bias and uncertainty much more likely in model output such as spawning stock biomass.

Bias in spawning stock biomass estimation was generally low for estimating models that assumed the correct source of process error when there was lower observation error. Reliable estimation of stock-recruit relationship parameters only appears possible in ideal situations with lower observation errors in age composition and indices, lower variability in recruitment process errors and large contrast in spawning biomass over time. We found little evidence of bias for many OM process error assumptions when there was contrast in fishing pressure even when there was greater observation error although it can lead to more variable estimates of natural mortality. For OMs where there was bias in natural mortality due to high observation error, estimating the stock-recruit relationship seemed to remove the bias. However, estimation of natural mortality can cause large differences between the true and estimated SSB (that may be unbiased on average) when there is less contrast in fishing pressure over time and higher observation error.

Retrospective patterns were generally weak for all estimation models regardless of the true source of process error, but they can be expected for recruitment even for the correct process error assumptions when observation error is high. When models did exhibit some retrospective pattern, estimating the mean natural mortality rate tended to remove it.

WP 5: Miller T. J., Applegate, A., Britten, G., Brooks, E. N., Fay, G., Hansell, A., Legault, C. M., Muffley, B., Wiedenmann, J. Factors affecting estimation of environmental effects on natural mortality

We completed a large scale simulation study with 288 operating models, each with 100 simulated data sets, and 12 estimated models fit to each simulated data set. The factors defining operating model configuration included the source of process error on the population (recruitment, recruitment and survival, recruitment and natural mortality), the degree of temporal variation and autocorrelation of the environmental covariate, the uncertainty in the observation of the covariate, the uncertainty in indices and age composition observations, fishing history, and the magnitude of the effect of the covariate on natural mortality. The estimating models make alternative assumptions on whether to include the environmental effect, whether the mean/intercept log natural mortality (log(0.2)) is estimated or known, and whether process errors are on just recruitment, recruitment and survival, or recruitment and natural mortality.

We found convergence of all estimation models was generally best when operating models assumed process errors in recruitment and survival, constant fishing rate, greater contrast in the true environmental covariate, and lower uncertainty in corresponding observations. Reliable convergence of all estimating models also occurred with the same process errors in the operating model and a step-change in fishing, but this also required lower uncertainty in index and age composition observations. Estimating models with process errors on recruitment and survival were unlikely to converge when the process errors in the operating model did not match whereas estimating models with process errors in recruitment and natural mortality converged for operating models without this match in certain cases. Probability of convergence generally decreased when the mean natural mortality rate parameter was estimated.

Whether the mean log-natural mortality was estimated or not, the best accuracy of AIC for model selection occurred for models with process errors on recruitment and survival. AIC accuracy was poor for models with process errors on recruitment and natural mortality. Estimating the mean natural mortality rate had small effects on the accuracy of AIC in selecting the appropriate process error. Estimating the mean log-natural mortality resulted in a small decrease in AIC accuracy for including the environmental effect. AIC was conservative for determining whether the environmental covariate affected natural mortality. AIC was very accurate in determining no effect when there was no effect in the operating model, but AIC often ranked the null model best when there was an effect. Accuracy of AIC for covariate effects improved with increased effect size, increased temporal contrast in the covariate, and lower uncertainty in observations.

We found no evidence of bias in estimation of environmental effects regardless of process error assumptions when there was low uncertainty in the environmental observations and large contrast in the environmental covariate. In most cases the relative error of the estimated environmental effect did not depend on the source of process error assumed in the estimating model. The worst bias was observed when OMs assumed R+S process errors, high uncertainty in covariate observations, low variability in the covariate, and low uncertainty in index and age composition observations. Simultaneously estimating the mean/intercept log natural mortality resulted in larger variation in the relative errors of the estimated environmental effect. Estimation of the intercept was reliable for all EM process error assumptions when the operating models assumed process errors on recruitment and natural mortality, contrast in fishing pressure over time, and lower observation error. Estimating the mean/intercept for log natural mortality generally resulted in highly variable estimates of annual natural mortality and spawning biomass and evidence of bias for some operating and estimation model assumptions about process error source. Again reliability of annual natural mortality estimates was generally improved with lower observation error uncertainty and contrast in fishing pressure.

Reliable detection of covariate effects requires informative data. AIC preferred simpler models than the true model when information content in data and contrast in covariates and abundance were low. The null model for environmental covariate effects (no covariate effect) was selected when contrast in the time series was low and/or uncertainty in observations was high. The selection of the null model by AIC also likely decreases with strength of the effect of the covariate on M. Similarly, when there was process error in recruitment and natural mortality, estimation models with process error only in recruitment were preferred presumably due to low variation in simulated natural mortality process errors. Covariate effect estimation can be robust to process error assumptions with high contrast in covariate and low observation error.

WP 6: Miller, T. J.. The nature of differences between analytic and projection-based equilibrium biomass reference points.

Equilibrium biomass and harvest reference points can be determined from assessment output either using analytic methods or projections of the population for a sufficient number of years. These projections can be either deterministic or incorporate stochasticity of the recruitment and other process errors. Central tendencies of the stochastic projection for equilibrium SSB or harvest will not be equivalent to the values from analytic methods. The difference is a result of summing log-normal random variables across ages for the stochastic projection approach and inability to accurately estimate the central tendency of this resulting distribution. For models without stock-recruit functions, contributions to SSB for each age less than the plus group can be accurately estimated, but the plus group cannot because it is also a sum of lognormal random variables. All age-specific contributions are expected to be biased when a stock-recruit function is used because recruitment is then a function of SSB. Analogous bias in equilibrium harvest reference points is expected because it is a similar function of abundance at age. The bias would exist for either SPR- or MSY-based reference points because the same SSB/R or Y/R calculations are used with equilibrium recruitment. Using The bias.correct option to TMB::sdreport will improve the estimate of the biomass reference point, but the improvement becomes negligible as the marginal variance of the autoregressive processes increase. Furthermore, the standard error of the estimated biomass reference point is not reported.

WP 7: Monnahan, C. C., and Correa, G. M. Estimability of time-varying growth in state space stock assessments.

State space models (SSM) are widely seen as best practices for fisheries stock assessment for processes like growth, but their immense flexibility means their statistical behavior can be unpredictable. We hypothesize that initial growth is driven by some external environmental signal, but the signal is observed with noise or not at all. Our narrow aim is to explore and contrast the statistical behavior and reliability of SSMs when initial growth is misspecified. We used a simulation testing framework with 6 operating models (OMs) which varied in the quantity of data (low vs high) and the size of the effect (beta) of an environmental driver on size at age 1 (L1; beta=0, 0.2, 0.4). 100 replicate data sets simulated from these were fitted to three estimation models (EM) which estimated constant growth, and variation in L1 driven either by an estimated environmental time series (matching the truth) or as an AR1 process. We found that when the OM matched the EM, the estimates of growth were unbiased for all parameters and had small relative errors. Estimation of mean L1 and k were poor for scenarios with a large beta and a mismatched EM, particularly the constant EM but also the AR1 EM. Finally, growth parameters were well estimated for all EMs when the OM had no time variation in L1. Our results suggest that using an AR1 on L1 is better than assuming constant, but notably worse than a model driven by the Ecov series. Furthermore, adding time variation to a model when none exists may not negatively affect model performance. Studies like this that examine the statistical behavior of SSMs in simplified simulation situations help build practical guidelines for advice on how to model time variation in growth in SSMs

WP 8: Hansell, A. Georges Bank Winter Flounder: Virtual Population Analysis to State-space

Historically, Georges Bank (GB) winter flounder has been assessed using a virtual population analysis (VPA). The last time this stock was assessed was the 2022 fall Management Track and the stock is currently in a rebuilding plan with a target date of 2029. The VPA has been plagued with uncertainty and diagnostic issues over the last several assessments including major retrospective patterns are poor cohort tracking. Research recommendations from the previous two assessments have been to explore a state-space framework for the stock. Additionally, the review panel from the last assessment recommended exploring different recruitment assumptions for projections. Here, we explore fitting the 2022 assessment inputs that were used for the last VPA to the Woods Hole Assessment Model (WHAM). We document the benefits of moving to WHAM and we propose a WHAM run for use in the 2025 Management Track assessment. The proposed run uses logistic normal distribution on all age composition and uses 2dar1 process on numbers at age. The proposed run has improved diagnostics compared to previous VPA runs and similar reference points and projections.

WP 9: Linton, B. Acadian redfish

The most recent Acadian redfish (*Sebastes fasciatus*) management track assessment occurred in 2023, using an Age-Structured Assessment Program (ASAP) model updated with data through 2022. While the ASAP model diagnostics have generally been considered good, the model failed to fit the spring and fall bottom trawl survey (BTS) indices at the end of the assessment time series, raising concerns with the peer review panels for the 2020 and 2023 management track assessments. A Woods Hole Assessment Model (WHAM) configuration was developed for Acadian redfish, with the short term goal of replicating the 2023 management track ASAP model results, and the long term goal of improving the model fit to the survey indices in a future management track or research track assessment. Six candidate WHAM configurations were explored during this research track. Due to the tight constraints on key model parameters (i.e., numbers-at-age in the first year and recruitment) in the 2023 ASAP model, the initial WHAM configuration was set up with similar tight constraints on model parameters. Those constraints were incrementally loosened in the successive five model configurations. The preferred WHAM configuration (Model 5) estimates equilibrium NAA in the first year, Beverton-Holt recruitment with annual deviations treated as i.i.d. random effects, and logistic selectivity functions for the fishery and the fall and spring BTS indices. The model fits to total catch, catch-at-age, fall BTS index, fall BTS age composition, spring BTS index, and spring BTS age composition were similar between Model 5 and the 2023 ASAP model. Estimated trends in F, SSB, and R were similar between Model 5 and the 2023 ASAP model. The F_{MSY}-proxy of $F_{50\%}$ is identical for Model 5 and the 2023 ASAP model. The SSB at the F_{MSY}-proxy is higher for Model 5 compared to the 2023 ASAP model. Projected catch and SSB are of similar magnitude between Model 5 and the 2023 ASAP model.

WP 10: Curti, K. and Hansell, A. Atlantic Mackerel

Atlantic Mackerel is currently assessed using ASAP. A WHAM model was developed here to (WP 10) following research recommendations from the 2023 spring Management Track

Assessment. The goal of the WHAM model was to improve diagnostics and retrospective patterns. The most recent ASAP model incorporated ages 1-10+, fishery and survey data from 1968-2022, and three index time series (rangewide egg index, NEFSC spring bottom trawl survey with the Albatross and Bigelow years treated as separate time series). A WHAM run configured the same way as ASAP successfully converged and produced similar trends as ASAP. Runs exploring the different age-compositions supported using a logistic normal age distribution that used an ar1 process to estimate zero values. Incorporating random effects into recruitment alone did not substantially improve model diagnostics so they were not included. Models with random effects on fleet selectivity converged but did not significantly improve diagnostics.   Full state-space models converged and including 2dar1 random effects on numbers-at-age led to improved Mohn's rho values. Thus, the selected model had similar inputs to the original ASAP model; however, it had logistic-normal distributions with an ar1 process on all age composition likelihoods and random effects using a 2dar1 process on numbers-at-age. The selected WHAM run had similar diagnostics as ASAP, improved retrospective patterns and similar temporal trends; however the estimates of uncertainty in recruitment, SSB, F and therefore, stock status, were larger. Similarly, short-term projections at $F_{msy\ proxy}$ exhibited the similar temporal trends as those from the ASAP model, but had much larger confidence intervals associated with the projected estimates. Overall, the proposed WHAM run produced similar trends as the previous ASAP model. There are several advantages from moving to WHAM: 1) the logistic-normal age compositions are self weighting; 2) including 2dar1 random effects on numbers-at-age reduces retrospective patterns; 3) the 2dar1 random effects on numbers-at-age also helps to account for recruitment correlations. However, the proposed WHAM run also has increased uncertainty around key parameter estimates. More work is needed to further determine if these uncertainties are realistic or an artifact of model specification. The base WHAM model that was configured like ASAP (fixed effects only) had similar trends to ASAP and similar estimates of uncertainty. At a minimum, moving over to WHAM and configuring it the same as ASAP would allow for future model development.

WP 11: Perretti, C. Gulf of Maine haddock WHAM case study

Currently, Gulf of Maine haddock is assessed using ASAP. Although the model diagnostics have been considered good overall, it has struggled to fit the rapid rise and fall in the observed survey indices in the last decade. In particular, the high abundance years immediately following the 2013 year class, as well as the most recent decline in abundance as that year class has aged. Similarly, the model developed a major retrospective pattern in the 2019 update due to the rapid increase in SSB implied by the survey observations. In the most recent update, a moderate retrospective pattern has developed in the opposite direction as SSB has declined rapidly. Possible drivers of these diagnostic problems include movement across stock boundaries, time-varying natural mortality, catch estimation errors, or some combination of the three. Previous work has not provided a definitive answer on the relative importance of each driver. A Woods Hole Assessment Model (WHAM) for GOM haddock was developed, with the main goal of reducing the retrospective error found in the most recent assessment and improving the fit to the survey indices in the terminal decade. The ASAP-like WHAM had large residuals in the indices in the terminal year, high retrospective error for recruitment, and inaccurate predictions in the terminal year. Therefore, three additional WHAMs were explored, all of which allowed for 2dar1 random effects (RE) on all ages. Model fit was generally similar across the RE WHAMs, with generally good diagnostics, an improved fit to the terminal decade of the indices compared to the ASAP model, and lower retrospective error than the ASAP model. The RE Dirichlet model was the only model to correctly predict the direction of change in the survey indices, and had the lowest simulation error of all RE models. Therefore, I recommend the RE Dirichlet model as the final model for GOM haddock. Trends in SSB, F, and recruitment are similar between the RE Dirichlet model and the accepted ASAP model, with generally lower F and higher SSB in the RE Dirichlet model. The overfishing reference point is similar between the two models with RE Dirichlet having a slightly lower Fmsy. SSBmsy is higher in the RE Dirichlet model due to the higher scale of recruitment throughout the timeseries. Projections are similar overall, with slightly lower projections in the RE Dirichlet model in the immediate future.

WP 12: Liljestrand, E. M., Bence, J. R., Deroba, J. J. The Effect of Process Variability and Data Quality on Performance of a State-space Stock Assessment Model (also in review)

State-space modeling is an emerging approach to age structured fisheries stock assessment that can accommodate multiple sources of variability in processes like recruitment, abundance, and selectivity. By maximizing the marginal likelihood by treating yearly deviations as random effects and then integrating them from the likelihood, these models can estimate multiple process variances. Several fisheries software packages have been developed that use a state-space framework with marginal likelihood, which has increased their popularity and usage across the U.S. Atlantic coast, Canada, and Europe. However, robust testing is still needed to gauge the applicability of these models and understand how they perform under a range of realistic variability in the process or observation error. Using an assessment model fit to Gulf of Maine Haddock as a baseline, we used a simulation-estimation procedure to test if state-space stock assessment models could produce unbiased and precise estimates over a range of process variances that extended from zero to well above the levels estimated in the Gulf of Maine Haddock assessment, or when observations were noisier (i.e., more variable around their true value) than had been assumed in the assessment. We fit alternative estimation models that differed in which processes errors were included (of recruitment, expected survival, and fishery selectivity). State-space models which specify random effects in all three processes produced unbiased and precise estimates of biomass and exploitation under most simulation scenarios and therefore are recommended except when variability in expected survival is absent (or very low), in which case the model is unlikely to converge. A conventional statistical-catch-at-age model with recruitment estimated as a fixed effect for each year, deterministic expected survival, and constant selectivity produced estimates that were comparable to the best state-space model, but do not provide internal estimates of process variances and did not perform well when recruitment was highly variable. This work will facilitate the use of state-space stock assessment models and choosing the parameterization that will produce the most accurate output to inform future predictions and management.