Common model diagnostics for fish stock assessments in the United States

Melissa A. Karp, Peter Kuriyama, Kristan Blackhart, Jon Brodziak, Felipe Carvalho, Kiersten Curti, E.J. Dick, Dana Hanselman, Daniel Hennen, James Ianelli, Skyler Sagarese, Kyle Shertzer, and Ian Taylor

U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Marine Fisheries Service

NOAA Technical Memorandum NMFS-F/SPO-240A
December 2022

Common model diagnostics for fish stock assessments in the United States

Melissa A. Karp, Peter Kuriyama, Kristan Blackhart, Jon Brodziak, Felipe Carvalho, Kiersten Curti, E.J. Dick, Daniel Hennen, James Ianelli, Skyler Sagarese, Kyle Shertzer, and Ian Taylor

NOAA Technical Memorandum NMFS-F/SPO240A

December 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Oceanic and Atmospheric Administration
Richard W. Spinrad, NOAA Administrator

National Marine Fisheries Service
Janet Coit, Assistant Administrator for Fisheries

Recommended citation:

Karp, Melissa A., Peter Kuriyama, Kristan Blackhart, Jon Brodziak, Felipe Carvalho, Kiersten Curti, E.J. Dick, Dana Hanselman, Daniel Hennen, James Ianelli, Skyler Sagarese, Kyle Shertzer, and Ian Taylor. 2022. Common model diagnostics for fish stock assessments in the United States. NOAA Tech. Memo. NMFS-F/SPO-240A, 27 p.

An earlier version of this report, F/SPO-240, was published online in December 2022. This revised version includes updates to the *Model Convergence* section, primarily in the sub-sections on *Second order conditions* and *MCMC convergence (for Bayesian methods only)*.

This report is available online at:

http://spo.nmfs.noaa.gov/tech-memos/

Table of Contents

ACRONYMS LIST

ABC	Acceptable Biological Catch
ADMB	Automatic Differentiation Model Builder
AFSC	Alaska Fisheries Science Center
AIC	Akaike Information Criterion
AMAK	Assessment Model for Alaska
ASAP	Age-Structured Assessment Program
ASMFC	Atlantic States Marine Fisheries Commission
ASPIC	A Surplus-Production Model Incorporating Covariates
BAM	Beaufort Assessment Model
BSAI	Bering Sea and Aleutian Islands
CASA	Catch-At-Size Assessment
CFMC	Caribbean Fishery Management Council
CIE	Center for Independent Experts
CNMI	U.S. Territories of American Samoa, Commonwealth of Northern Mariana Islands
CPUE	catch per unit effort
FMP	Fishery Management Plan
GMACs	Generalized Model for Alaska Crabs
GMFMC	Gulf of Mexico Fishery Management Coucnil
GOA	Gulf of Alaska
HMS	Highly Migratory Species
HPD	Highest Posterior Density
ICCAT	International Commission for the Conservation of Atlantic Tunas
ICES	International Council for the Exploration of the Sea
INLA	Integrated Nested Laplace Approximation
ISC	International Scientific Committee
JABBA	Just Annother Bayesian Biomass Assessment
MAFMC	Mid-Atlantic Fishery Management Council
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
NEFMC	Northeast Fishery Management Council
NEFSC	Northeast Fisheries Science Center
NPFMC	North Pacific Fishery Management Council
NUTS	No U-Turn Sampler
NWFSC	Northwest Fisheries Science Center
OFL	Overfishing Limit
PFMC	Pacific Fishery Management Council
PIFSC	Pacific Islands Fisheries Science Center
SAFMC	South Atlantic Fishery Managemetn Council
SEDAR	SouthEast Data Assessment and Review
SEFSC	Southeast Fisheries Science Center
SS3	Stock Synthesis 3

SSC	Scientific and Statistical Committee
STAR	Stock Assessment Review Panel
SWFSC	Southwest Fisheries Science Center
WCPFC	Western and Central Pacific Fisheries Commission
WPFMC	Western Pacific Fishery Management Council

1. INTRODUCTION

Fishery stock assessments are mathematical models that estimate the current status of a fish stock and are the foundation of sustainable fisheries management. Stock assessments typically rely on a number of data types to estimate key biological processes, the effects of fishing activity, and, in some cases, the influence of environmental conditions on the past and current stock status. Each stock assessment is unique and there is no objective criterion that simultaneously summarizes the results of an assessment, determines if the model fits the data adequately, and evaluates whether the model is well-specified. As a result, stock assessment analysts rely on a suite of model diagnostics to ultimately arrive at a single base model (or in some cases a suite of base models termed an ensemble of models) to inform management decisions for a specific fish stock.

There are a number of regional differences in developing stock assessment models, and our goal here is to document the model diagnostics commonly used to develop a base model, or ensemble of models, for stocks managed by NOAA Fisheries, while incorporating the regional variation in methods. Models, modeling software, and diagnostic methods are tailored based on data availability, species life history, and unique fishery conditions, all of which vary from stock to stock and region to region. We note that some model diagnostics are broadly applicable to all regions of the country, whereas others have more limited use or may be applied differently across regions.

We do not claim that this document is comprehensive nor do we wish for it to impede future development of model diagnostic methods or implementation of diagnostic best practices. This document provides information on diagnostic methods for NOAA Fisheries stock assessment scientists, especially those new to the agency, as well as management partners from the Fishery Management Councils and other assessment stakeholders. However, we note that this a highly technical document and point readers to additional stock assessment and quantitative modeling resources such as Quinn and Deriso (1999) and Haddon (2011) to help understand and provide additional context and background on the various methods described herein. We hope also to encourage continued discussion of best practices and to clarify technical concepts that often arise in the stock assessment review process.

2. U.S. REGIONAL CONTEXT

In this section we provide an overview of the stock assessments conducted in each region, including the types of stocks assessed (life histories), kinds of data used (fishery-independent survey/fishery-dependent survey), time frames of data, and software packages. A discussion of the assessment review process in each region is beyond the scope of this document, and we recommend readers go to Lynch et al. (2018) for a full overview of the regional assessment review process. However, we do note that in all regions stock assessments are subject to peer review at different levels depending on the number of new methods applied, with more change resulting in more thorough reviews. Differences between regions highlighted in this section may help readers understand the various applications of the common model diagnostics across regions.

2.1. Northwest Fisheries Science Center

The Northwest Fisheries Science Center (NWFSC) assesses groundfish (numerous rockfishes and flatfishes along with Pacific hake, lingcod, sablefish, and others) and Pacific salmon. The Southwest Fisheries Science Center (SWFSC, discussed next) also assesses groundfish and Pacific salmon; however,

for both the NWFSC and SWFSC, assessments of Pacific salmon follow a unique framework[1] that differs substantially from many marine fish assessments and are not addressed in this document. The groundfish assessments conducted by NWFSC and SWFSC differ only in the specific biology, data sources, and needs for each species. Division of groundfish assessments between the two science centers depends on staff availability and expertise with data sources (e.g., stocks that are more abundant in California are more likely to be assessed by SWFSC staff).

Groundfish assessments rely on both fishery-dependent and fishery-independent data sources. Fishery-dependent data include annual landings, discard (rates, numbers, or tons), catch per unit effort (CPUE), and age and size compositions. Catch reconstructions for groundfish often extend back to the late 1800s or early 1900s and are assumed to cover the entire history of the fishery (models begin from an unfished state). CPUE time series and composition data are typically from more recent years (1980-present), with some exceptions. Fishery-independent data include time series of relative abundance from trawl and hook-and-line surveys, indices of recruitment (SWFSC Rockfish Recruitment and Ecosystem Assessment Survey, 1983-present), absolute abundance from visual surveys (2002 and 2012), and time series of spawning output (CalCOFI, 1951-present).

Assessments are conducted with integrated statistical catch-at-age models using Stock Synthesis 3 (SS3), a widely-used and flexible software for conducting assessments. The science is reviewed as part of the Pacific Fishery Management Council (PFMC) process by the Stock Assessment Review Panel (STAR) and the Scientific and Statistical Committee (SSC), and if accepted as best scientific information available, adopted by the Pacific Fishery Management Council.

2.2. Southwest Fisheries Science Center

The Southwest Fisheries Science Center (SWFSC) assesses highly migratory species (HMS; Pacific bluefin tuna, albacore, thresher shark, billfish) and coastal pelagic species (Pacific mackerel, Pacific sardine, northern anchovy), in addition to groundfish (numerous rockfishes and flatfishes), and Pacific salmon as described in the NWFSC section above.

Highly migratory species assessments rely on fishery-dependent catch, fishery-dependent abundance indices, and size compositions. These assessments often contain data from many fishing fleets (~10-30) that utilize pole-and-line gear and longlines in areas of the northwest and northeast Pacific Ocean. The modeling time frames can be relatively long (e.g., Pacific bluefin models begin in the early 1950s) to relatively short (e.g., albacore models begin in the mid-1990s). The assessments are conducted as part of international working groups with scientists from the Pacific Islands Fisheries Science Center (PIFSC) and agencies in Japan, Korea, and Chinese Taipei as part of the International Scientific Committee (ISC)[2].

Coastal pelagic species rely on the fishery-independent acoustic-trawl survey which spans the west coast of the U.S., fishery-dependent catch, and age and size compositions. Fishery catch and composition data, for Pacific sardine specifically, date back to the early 1980s. However, the assessment time frames are generally quite short as generation time for these species are less than ten years. For example, the Pacific sardine and Pacific mackerel plus group begins at age 8. Northern anchovy are rarely observed older than 4 years old. Short modeling time periods cover multiple generations as the life histories of coastal pelagic

[1] https://www.pcouncil.org/salmon-document-library/
[2] https://isc.fra.go.jp/

species are shorter than those for highly migratory species and groundfish. Management decisions are based on biomass forecast for the upcoming fishing year, rather than relative to reference points.

Assessments are conducted with integrated statistical catch-at-age models using SS3. Groundfish and coastal pelagic species assessments are reviewed as part of the Pacific Fishery Management Council process by a STAR, typically composed of a subset of members from the SSC and independent reviewers from the Center for Independent Experts (CIE). If deemed suitable, the stock assessment is then presented to the full SSC and for consideration as best scientific information available. The voting members of the PFMC then adopt catch guidelines based on the scientific and management uncertainties.

2.3. Alaska Fisheries Science Center

The Alaska Fisheries Science Center (AFSC) is responsible for groundfish, crab, and scallop stock assessments for stocks found in the Bering Sea, Aleutian Islands, and Gulf of Alaska. There are two groundfish fishery management plans (FMPs), the Bering Sea and Aleutian Islands (BSAI) and the Gulf of Alaska (GOA); one crab FMP (BSAI only); and one scallop FMP.

AFSC and Alaska Department of Fish and Game have multiple long time series of fishery-independent data streams for abundance and demographic data including long-line, bottom trawl, and pot surveys. The earliest consistent time series begin in 1979 (AFSC longline) and 1982 (Eastern Bering Sea bottom trawl). Most stock assessments use reconstructed catch estimates starting between 1960 and 1977, with a full domestic observer program providing reliable catch estimates (landings plus discards) starting in 1991. The observer program provides ages, lengths, and weights of the catches and fishery-dependent CPUE in a few cases.

Groundfish, crab, and scallop FMPs are managed under tiers roughly corresponding to data availability and reliability. The scallop assessment focuses only on weathervane scallop and uses a fixed acceptable biological catch (ABC)/overfishing limit (OFL) based on average catch for federal management, and the State of Alaska sets guideline harvest levels. The majority of targeted and high abundance crab and groundfish stocks are assessed using statistical catch-at-age/length models (e.g., Assessment Model for Alaska (AMAK), SS3, and Generalized Model for Alaskan Crabs (GMACS)), while the non-target stocks are assessed using biomass-based methods (time-series models and survey averages) or catch-only models (average or maximum catch). Groundfish life histories range from fast growing higher mortality gadids ($M \approx 0.4$) to long-lived, slow growing rockfish such as rougheye rockfish ($M=0.03$). Crabs are not easily aged so length based models are applied and present unique challenges such as estimating terminal molt probability. The assessment review process comes from the North Pacific Fishery Management Council (NPFMC) Plan Teams and SSCs.

2.4. Pacific Islands Fisheries Science Center

The Pacific Islands Fisheries Science Center (PIFSC) works on stock assessments for a diverse set of fishery resources. These are insular and pelagic species such as shallow-water reef fishes; deep-water bottomfishes including snappers and groupers; small pelagics; and HMS including billfishes, sharks, and tunas. The assessments are conducted through collaboration between NOAA Fisheries and state, territorial, and international fisheries agencies and scientists. The stock assessments are used by national and international fisheries management bodies, including the Western Pacific Regional Fishery Management Council (WPFMC) and the Western and Central Pacific Fisheries Commission (WCPFC).

Stock assessments of insular species are conducted for the Hawai´ian Archipelago and the U.S. Pacific Territories. Reef fish assessments have typically been conducted with length-based methods and have used fishery-independent data from diver surveys for monitoring coral reef ecosystems since the 2000s. The bottomfish resources of the main Hawai´ian Islands are primarily assessed using Bayesian production modeling approaches (i.e., JABBA, Just-Another Bayesian Biomass Assessment) with catch time series going back to the late-1940s and a fishery-independent camera and deep-longline survey since the 2010s. In recent years, an integrated statistical catch-at-age model has been developed for uku, or the blue-green Hawai´ian snapper, using SS3. The bottomfish resources of the U.S. Territories of American Samoa, Commonwealth of Northern Mariana Islands (CNMI), and Guam have also been assessed using production modeling approaches with catch and CPUE time series going back to the 1980s. Stock assessments conducted by PIFSC for insular domestic species are reviewed in the Western Pacific Stock Assessment Review process[3]. This includes assessments for fishery resources in the Hawai´ian Archipelago and U.S. territories of American Samoa, the Commonwealth of the Northern Mariana Islands and Guam.

The stock assessments of highly migratory billfish, pelagic sharks, and tunas are developed as part of the International Scientific Commission for Tuna and Tuna-Like Species (ISC) in the North Pacific as well as the WCPFC. The assessments of highly migratory pelagic species are typically based in national and international waters and are conducted in collaboration with international scientific working groups of the ISC or the Secretariat of the Pacific Community[4] . These assessments typically apply standardized CPUE from multiple fleets with size composition data in an integrated assessment modeling framework such as SS3 or Multifan-CL. The fishery-dependent data sources from various countries contribute data on catch time series, including discards for sharks where possible. Time frames for HMS vary with data quality; some assessment time frames begin in the 1950s while others start in the 1990s. Key data sources for these assessments include relative abundance indices and size compositions from commercial longline, purse seine, and other fisheries. Stock assessments for some HMS are based on one best-fitting statistical catch-at-age model (one base model), while for others (e.g., tropical tunas and more recently for Pacific blue marlin), ensemble model approaches are typically applied to account for uncertainty in the assessment model structure. International stock assessments conducted for HMS in collaboration with the PIFSC are peer-reviewed by the ISC Plenary[5] and the WCPFC Scientific Committee[6] .

2.5. Northeast Fisheries Science Center

The Northeast Fisheries Science Center (NEFSC) assesses fisheries that occur north of Cape Hatteras, North Carolina. These include many stocks of groundfish, such as Atlantic cod, haddock, pollock, and various flatfish; pelagic species like Atlantic herring and mackerel; as well as invertebrates including bivalves, lobster, and squid. These stocks are managed under the auspices of two federal fishery management councils, the Northeast Fishery Management Council (NEFMC) and the Mid-Atlantic Fishery Management Council (MAFMC), depending on where the primary fisheries for each stock are based. Some stocks such as black sea bass, summer flounder, scup, bluefish, and Atlantic herring are co-managed with the states through the Atlantic States Marine Fisheries Commission (ASMFC), and some stocks (including eastern Georges Bank cod, haddock and yellowtail flounder) are also co-managed with Canada through the Transboundary Management Guidance Committee. NEFSC is also involved in the

[3] https://www.fisheries.noaa.gov/pacific-islands/population-assessments/western-pacific-stock-assessment-review
[4] https://www.spc.int/
[5] https://isc.fra.go.jp/meetings/future_meetings.html
[6] https://www.wcpfc.int/folder/scientific-committee

assessments of several state-managed anadromous stocks, such as striped bass, American shad, and river herring, whose management falls to ASFMC.

The data used at NEFSC are relatively historically rich, with both fishery-dependent and fishery-independent sources extending back several decades for most stocks. Catch time series often extend back to the early 1900s and bottom trawl surveys have been conducted in the region since the 1960s. Catch-at-age and size for fisheries and surveys have been reliably collected since the 1980s in many cases. Observer coverage in the northeast is considerably less than in some other regions, with generally less than 10% of commercial trips carrying an observer.

The primary assessment framework used for groundfish and pelagics is a statistical catch at age model, ASAP (Age-Structured Assessment Program). However, groundfish assessments in this region are beginning to transition to the state-space Woods Hole Assessment Model (WHAM), which allows for improved consideration of time-varying processes via random effects or environmental-productivity links within the model (Stock and Miller, 2021). Invertebrates typically use other platforms, including length-based models such as CASA (catch-at-size assessment) and SS3. Stock assessments are subject to peer review by the SSC of the NEFMC or MAFMC at different levels depending on the number of new methods applied, with more change resulting in more thorough reviews[7].

2.6. Southeast Fisheries Science Center

The Southeast Fisheries Science Center (SEFSC) assesses HMS (tunas, billfish, pelagic sharks, coastal sharks); coastal pelagic species (cobia and mackerels); snappers, groupers, tilefish, triggerfish, amberjack, and other reef fishes in the South Atlantic and Gulf of Mexico regions; menhaden in the Atlantic and Gulf of Mexico regions; data-limited reef fishes and invertebrates (e.g., Caribbean spiny lobster) in the U.S. Caribbean; and shrimp (pink, brown, and white). Stock assessments conducted by the SEFSC are used by three regional Fishery Management Councils (U.S. Caribbean (CFMC), Gulf of Mexico (GMFMC), South Atlantic (SAFMC)) and Interstate and International Fishery Commissions.

Stock assessments of HMS including tunas, billfish, and pelagic sharks are developed as part of the International Commission for the Conservation of Atlantic Tunas (ICCAT) process, although some HMS shark assessments for coastal species are developed and reviewed via the SouthEast Data Assessment and Review (SEDAR) process. Due to their transboundary ranges, HMS assessments rely primarily on fishery-dependent data from various countries that contribute data on catch (and discards for coastal sharks where available; discards are not reported to the ICCAT), relative abundance indices (sometimes as joint CPUE indices), and age and size compositions. Commercial fisheries (longlines, purse seines, etc.), and to a lesser extent recreational fisheries, harvest these species as targeted landings, with little data available concerning discards. Indices of relative abundance and size compositions are derived from fishery-independent surveys where available (e.g., coastal sharks). Stock assessments for HMS are generally conducted with integrated statistical catch-at-age models using SS3, although other approaches are frequently applied (e.g., production models such as JABBA or ASPIC (A Surplus-Production Model Incorporating Covariates)).

Stock assessments of coastal pelagic species rely primarily on fishery-dependent data sources to contribute data on catch, discards, relative abundance indices, and age and size compositions. Fishery-independent data sources are incorporated for mackerels but not for cobia due to the lack of adequate

[7] https://s3.amazonaws.com/nefmc.org/NRCC_Assessment_Process_Version-18Feb2022_508.pdf

sampling. Commercial (primarily line and/or gillnet gears) and recreational fisheries harvest these species as both targeted landings and as discards (primarily regulatory), with bycatch removals by the shrimp fishery being substantial. Indices of relative abundance and size compositions are derived from fishery-independent groundfish trawl surveys or plankton surveys for mackerels. Model time periods range from starting in 1886 for Spanish mackerel in the Gulf of Mexico to the late 1920s for both king mackerel and cobia in the Gulf of Mexico and to the 1980s for cobia in the Atlantic. Stock assessments for these species are primarily conducted with integrated statistical catch-at-age models using SS3 or the Beaufort Assessment Model (BAM; used for South Atlantic Spanish mackerel and cobia) and are developed during the SEDAR process. Models are reviewed by SSC's of both the GMFMC and SAFMC, and if accepted as best available science, adopted by each Council.

Stock assessments of reef fishes in the South Atlantic, such as groupers, snappers, porgies, triggerfish, sea bass, grunts, and tilefishes, utilize data on landings, discards, indices of abundance, size compositions, and age compositions. Because these stocks comprise multi-species fisheries, discard rates can be high, with discard mortality representing a substantial portion of removals. For many of these reef fishes, recreational fleets (headboats, charter boats, and private anglers) capture more fish than commercial fleets (handline, longline, traps, and diving). The initial year included in the model for South Atlantic assessments varies depending on the species, ranging from 1950 to the 1980s. Length compositions of landings are generally available starting in the 1970s or 1980s, and length compositions of discards are available starting in the 2000s. Age compositions of landings are generally available starting in the 1990s. Most assessments utilize fishery-dependent CPUEs, especially one developed from commercial handline logbook data (starting in 1993) and one developed from headboat logbook data (starting in the late 1970s or early 1980s). The primary fishery-independent indices are developed from chevron trap sampling (starting in 1990) and from video sampling (starting in 2011). Because sampling by these two gears is not independent (cameras attached to traps), the two indices are often combined prior to being used as stock assessment input. Stock assessments in the South Atlantic are conducted with integrated statistical catch-at-age models using the BAM and are developed via the SEDAR process. Models are reviewed by the SAFMC's SSC, and if accepted as best scientific information available, adopted by the Council.

Stock assessments of reef fishes in the Gulf of Mexico rely on numerous fishery-dependent and fishery-independent data sources to contribute data on catch, discards, relative abundance indices, size compositions, and age compositions. Commercial (primarily line and longline gears) and recreational fisheries (headboats, charter boats, private or shore anglers) harvest reef fish as both targeted landings and as discards (generally regulatory in nature, but can be quite high for some stocks due to either magnitude or high discard mortality rates). Bycatch removals of some species (e.g., red snapper) by the shrimp fishery can be substantial, and severe red tide events have been incorporated into grouper stock assessments as a source of episodic mortality. Composition data include retained length compositions (since the mid-1980s), discarded length compositions (since the mid-2000s via NOAA Fisheries and Florida Fish and Wildlife Conservation Commission Observer Programs), and retained age compositions (since the early 1990s). Fishery-independent bottom longline surveys, groundfish trawl surveys, and video surveys are important data streams for many species and index both abundance and size composition (of adults, juveniles, or both). The modeling time frames vary considerably from long time series (e.g., 1880 start for red snapper) to shorter time series (e.g., 1986 for red grouper and scamp). Stock assessments in the Gulf of Mexico are conducted with SS3 and are developed via the SEDAR process. Models are reviewed by the GMFMC's SSC and, if accepted as best available science, adopted by the Council.

Stock assessments of menhaden in both the Atlantic and Gulf of Mexico regions (two distinct species) are conducted using fishery-dependent data (purse seine fisheries) and fishery-independent data (inshore state surveys) sources. Stock assessments are conducted using BAM in each region, with assessments developed and reviewed during the SEDAR process. The Gulf and Atlantic models are utilized by the Gulf States Marine Fisheries Commission and ASMFC, respectively, both of which are responsible for setting annual catch advice.

U.S. Caribbean stock assessments are conducted separately for three island units (Puerto Rico, St. Thomas/St. John, and St. Croix) and are data-limited in nature because of variable data quantity and quality in each region. Assessments generally rely on fishery-dependent data sources including catch (and discards if available), CPUE indices, and size compositions. Dominant commercial and recreational fisheries differ considerably between island units and employ a wide range of gears (e.g., handlines, traps, diving, etc.) to harvest reef fish and invertebrates. Size data of fish retained by the commercial fisheries are collected by port samplers, but data quality and quantity may vary between regions and years. Recreational data are only sparsely available for Puerto Rico (catch and discards only, no size information) and are completely lacking for the other two island units. Although many fishery-independent data sources exist, there have been numerous caveats associated with their use in stock assessments. Stock assessments in the U.S. Caribbean have been conducted using data-limited approaches such as the non-equilibrium mean length estimator via the SEDAR process. Models are reviewed by the CFMC's SSC, and if consistent with best scientific information available, adopted by the Council. The most recent data-moderate integrated assessment conducted for the Caribbean spiny lobster was the first accepted for use in developing management advice for the region.

Stock assessments for pink, brown, and white shrimp in the Gulf of Mexico region are conducted using fishery-dependent data (trawl fisheries) and fishery-independent data (groundfish surveys) sources. Unique to shrimp, stock assessments are not reviewed during the SEDAR process. Penaeid shrimp stock assessments using SS3 have been vetted and reviewed by the GMFMC SSC and Special Shrimp SSC since their inception in 2009. Given the short lifespan of shrimp, analysts are considering other assessment approaches, such as empirical dynamic modeling.

3. COMMON MODEL DIAGNOSTICS

There are many diagnostics that are commonly used when evaluating a stock assessment model. In this section we provide an overview of the currently used model diagnostics. The diagnostics are grouped into categories based on the main topic they address. These categories are: model convergence, goodness-of-fit, and model robustness. For each diagnostic, we describe the goals, lay out key considerations, and provide examples of its use within existing stock assessments.

3.1. Model Convergence

Stock assessment models typically rely on a numerical optimization procedure to estimate parameters. Convergence indicates that the optimization procedure has found a minimum of the objective function (i.e., that a solution has been found), and diagnostics can help determine if that minimum is a global minimum (i.e., best solution). A lack of model convergence indicates that the optimization criteria used for statistical estimation are not satisfied. As a result, parameter estimates and management values should not be relied upon. A particular model may not converge due to poor model specification (e.g., over-parameterization), non-informative data, data conflicts (e.g., two indices of abundance have opposite trends), poor initial values, or insufficient iterations of the optimizer. Model convergence can generally be

classified into two types, that which provides the maximum likelihood (or highest posterior density) and that which provides adequate coverage to estimate the full posterior probability. The latter specifically refers to Bayesian applications where marginal distributions may matter. Convergence, either for optimization or posterior distribution integration (see MCMC below) indicates a stable, repeatable set of estimates.

Convergence should be checked using multiple diagnostics as no single test is sufficient. There are several useful diagnostic checks for evaluating the convergence of a model. The main tests include examining (1) if parameter estimates are near or at bounds, (2) if the final gradients are small in magnitude, (3) if the Hessian matrix is positive definite, and (4) if model estimates are robust to a range of random initial conditions (a.k.a. jitter analysis). Below we provide details on these main approaches but note other good practices include calculating the condition number of the covariance matrix or ratio of the largest to the smallest eigenvalue, which provides a measure of how well-conditioned the covariance matrix is. In this case, a very large condition number indicates that there is ill-conditioning or collinearity in the calculated covariance matrix, which in turn indicates that the parameter estimates are less reliably estimated. The covariance matrix should also be evaluated to determine if there is confounding of parameter estimates, which indicates likely over-parameterization of the model.

3.1.1. Checking for parameters estimated at bounds

Many modeling platforms allow for parameter estimates to be bounded by the analyst. For Bayesian stock assessments, these bounds would be defined as priors which may be informative or uninformative, where either can be based on outside analysis or a previous model. Bounding parameters is often done to prevent the optimization algorithm from searching extreme regions of the potential parameter space; however, when parameters are estimated at or close to these bounds, it indicates that the data do not inform estimation, that there are problems with the assumed model structure, or that a better fit may be found outside of the chosen parameter or prior bounds. Uncertainty calculations are unreliable when a parameter estimate is near or at a bound. Therefore, one of the first steps for evaluating model convergence is checking to see if any parameters are estimated at a bound. This analysis is commonly employed by all regions in the U.S.

Software used to conduct a stock assessment will often have a number of output files[8]. Typically, these files will report parameter estimates along with the lower and upper bounds. Checking which parameters are near bounds is fairly straightforward. However, resolving parameters on bounds can be a more significant undertaking, depending on why it occurs. In some cases, the issue can be resolved by fixing problematic parameters or by estimating them with informative priors; in other cases, the model might require modification (see section 3.3 on Model Consistency).

3.1.2. First order conditions: Checking that the final gradient is relatively small

Optimization of a nonlinear objective function for a stock assessment model generally requires the iterative calculation of numerical estimates of the parameter vector. The convergence of the iterative sequence of estimates needs to be evaluated to ensure that a possible solution has been obtained.

Here the convergence of a sequence of estimates to a solution is evaluated by either (1) measuring the distance of the model gradient relative to zero, or (2) measuring the relative or absolute distance between

[8] For ADMB-based models (such as SS3, AMAK, ASAP, and BAM), we note that the newest ADMB version 13.0 provides improved bounds checking relative to earlier versions.

successive estimates of the maximum likelihood estimate (MLE) parameter vector. The first approach involves calculating the gradient of the objective function for each iterative estimate of the solution, and verifying that the gradient is effectively equal to the zero vector at the final solution. Alternatively, if the distance between successive estimates of the MLE is close to zero, it implies that they are equal, and thus the algorithm being used to search parameter space cannot reduce the negative log-likelihood further. There is no standard cut-off for what is considered "close-to-zero", and regional differences exist in how this is determined, but the default used in SS is 0.0001, and thus could be used as a reasonable starting point. The two conditions described here are consistent with the existence of a stationary point, which could be a local minimum or a saddle point. Information from the Hessian matrix and jittered starting values described in sections 3.1.3 and 3.1.4 below can help distinguish between these types of stationary points.

The convergence metric used to stop the iterative gradient descent optimization approach may also be influential (e.g., Subbey 2018, see Caveat 3). As a result, it is recommended to check that the iterative solution is robust to the stopping criterion used, whether it is the relative or absolute distance between successive parameter vector estimates or the gradient being zero.

3.1.3. Second order conditions: Hessian matrix is positive definite

Optimization of nonlinear stock assessment models are indicated when the gradients of the objective function with respect to parameters are near zero and a second order condition is satisfied. The second order condition is satisfied when the covariance matrix (inverse Hessian matrix; matrix of second order partial derivatives with respect to the parameters) is positive definite at the proposed solution. When the Hessian is positive definite (e.g., all eigenvalues are positive) at the point estimate, it essentially means that the function is curving upward and you have found a local minimum. This check is employed by all Regions.

This condition can be checked by applying a Cholesky decomposition or an eigenvalue analysis (i.e., singular value decomposition) to the estimated Hessian matrix to verify that all of the eigenvalues are positive. For ADMB-based platforms, checks of the Hessian matrix are reported in the console output and in the various output files and if it fails, asymptotic approximations to parameter variances will be unavailable.

3.1.4. Alternative initial parameter values (i.e., jitter run)

Nonlinear models need starting values and sometimes it is not trivial to generate these values. Poor choices can lead to non-convergence or convergence upon a local rather than global minimum. It is important to evaluate whether the calculated solution that minimizes the nonlinear objective function is sensitive to the initial parameter values. Rerunning a model repeating starting estimation at different initial parameter values is termed model jittering. Jittering checks to see if any of the randomly generated starting values of parameters results in a better solution (i.e., smaller total negative log-likelihood) than the reference model. Jittering can thus be conducted to provide confidence that the model has converged upon a global solution rather than a local minimum, regardless of the starting values used (but see Subbey, 2018).

Starting the model using alternative but wisely chosen starting values can provide evidence that the model estimates are based on the global minimum if all alternatives lead to the same likelihood. The first step of this test is to change the initial values used for all estimated parameters and then refit the model. Typically, this is done 50-100 times. Technical details of the SS3 implementation of the jitter within parameter bounds can be found in the user manual (Methot et al., 2022). Next the analyst should

summarize the total likelihood values from each jitter run, and look at which likelihood components are changing, if any. If all runs converge upon a similar solution, consider whether the investigated range to select alternative initial parameter values is too narrow.

Jitter analyses are relevant and can be applied in any estimation framework used for a nonlinear model. Sensitivity of the calculated optimal solution to initial parameter values should be evaluated for frequentist assessment models to ensure that the numerical results are robust. For Bayesian assessment models, Markov Chain Monte Carlo (MCMC) simulations to sample the posterior should be conducted with multiple chains using different initial parameter vectors. Note that the jittering process should be applied enough times to check if a better solution can be found (i.e., the first solution is not at a local minimum). In complex assessments, one should check that if a new, better solution is achieved, the components of the likelihood that contributed to the finding are understood and plausible. Finally, such an exercise can avoid solutions in local minima, but there is no guarantee that the new solution is the global minimum. Here it is also important to note that the jitter analysis is a confirmatory analysis that supports but does not demonstrate that the putative MLE is a global minimum (e.g., Subbey 2018, see Caveats 2 and 4).

3.1.5. MCMC Convergence (for Bayesian methods only)

When using Bayesian approaches to sample from the posterior distribution via MCMC, some additional sets of diagnostic tests for convergence should be applied, such as visual inspection of trace plots and density plots of parameters across MCMC iterations. Trace plots provide a quick initial check of whether successive iterations appear to be independent and identically distributed random samples. The R package, "coda", provides these and gives some convergence tests (e.g., the Gelman Rubin, Geweke, and Heidelberger and Welch tests; Gelman et al., 2013). Example application of this approach is available for the assessment of the bottomfish complex (Langseth et al., 2018). In addition, some advanced MCMC algorithms and diagnostics are presently available for most assessment models. This involves an implementation of the Hamiltonian Monte Carlo approach with the "No U-Turn Sampler" (NUTS; for details please see Stan Development Group 2017, Carpenter et al. 2017). As implemented through the R package "adnuts" (Monnahan and Kristensen, 2018) this approach provides a fast, easily parallelized way to test models for posterior distribution convergence. The suite of diagnostics is interfaced with the tools available in Stan and includes the effective sample size, the ratio of variances within and among chains (Rhat), an interactive diagnostic tool ShinyStan (Gabry and Veen 2022), and quickly highlights issues with model specifications and parameterizations. A baseline convergence check could be ensuring that the maximum Rhat is small (<1.05) and that the minimum ESS is sufficient for inference, with hundreds of samples typically being sufficient for estimating means, but more required for tail probabilities. Edwards et al. (2022) used adnuts to investigate correlation between parameters and likelihood surfaces and check posterior estimates and marginal distributions with their asymptotic equivalents.

3.2. Goodness of Fit

Goodness of fit is a measure of how well the predictions from a statistical model match the data. A stock assessment model with a good fit to the data is more likely to account for important processes in the population and to produce useful estimates of stock size, status, productivity, and projections to inform management. The presence of systematic misfit to the data is a sign that the model may be misspecified. For a model to be useful, however, characteristics other than goodness of fit should also be considered (e.g., parsimony and out-of-sample predictive ability). Here we describe two approaches used to evaluate the goodness of fit of a model, residuals and likelihood profiles.

3.2.1. Residuals

Residuals represent the difference between the observation and the value predicted by the model. Residual analysis is probably the most common diagnostic used to identify a model's goodness-of-fit and is used across all regions in the U.S. Contemporary stock assessment models often include multiple data types (e.g., indices of abundance, size and age compositions, catch, discards), and given that residuals will be available for every type, evaluating tradeoffs in residuals is often a key component of developing and fitting models. For example, in the Gulf of Mexico red grouper stock assessment, there was poor fit to commercial vertical line fishery discard data (Fig. 1; top panel), but a better fit to the commercial longline fishery discard data (Fig. 1; bottom panel).

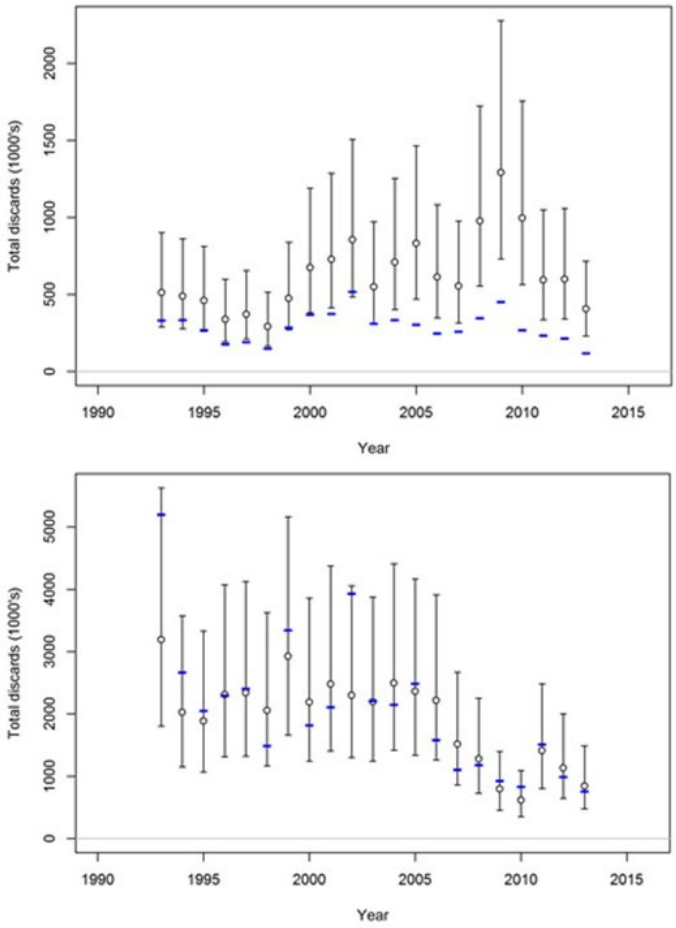

Figure 1: Observed (open circles) and predicted discards (blue dashes) in 1000's of fish of Gulf of Mexico red grouper from the commercial vertical (top panel) and longline (bottom panel) fleets, 1993-2013 (SEDAR, 2015).

Model fit can be evaluated by looking at the magnitude of the residuals relative to observation error or the presence of trends in residuals. Ideally, residuals should be minimized, randomly distributed (i.e., displaying no prominent patterns or systematic deviations), consistent with distributional assumptions (e.g., normally distributed), and have few outliers (i.e., observations that are significantly outside the range of model predictions). For example, residuals that are consistently positive or negative would indicate that an important process is not accounted for in the model. A runs test can reveal whether sequences of positive or negative residuals are longer than might be expected simply from chance (Wald and Wolfowitz, 1940; Carvahlo et al., 2021).

Raw residuals can reveal lack of fit, such as residual patterns, but there are more sophisticated diagnostics for residuals. The interpretation of raw residuals may be challenging for asymmetrical or discrete statistical distributions. Thus, multiple residual definitions have been developed, each with properties that increase their utility in assessing goodness of fit. Residuals that are scaled by a measure of variability can help identify observations that are improbable given the model. A common approach is to divide the raw residuals by an estimate of the residual standard deviation (sometimes referred to as 'standardized residuals'). This practice generates residuals that describe the distance between observed and predicted values in standard deviation units. Standardized residuals with an absolute value above a certain threshold (e.g., 2 or 3 standard deviations) can be identified as outliers. Residuals can also be plotted in more complex ways than just observed versus expected as a quick but coarse means of diagnosing residual

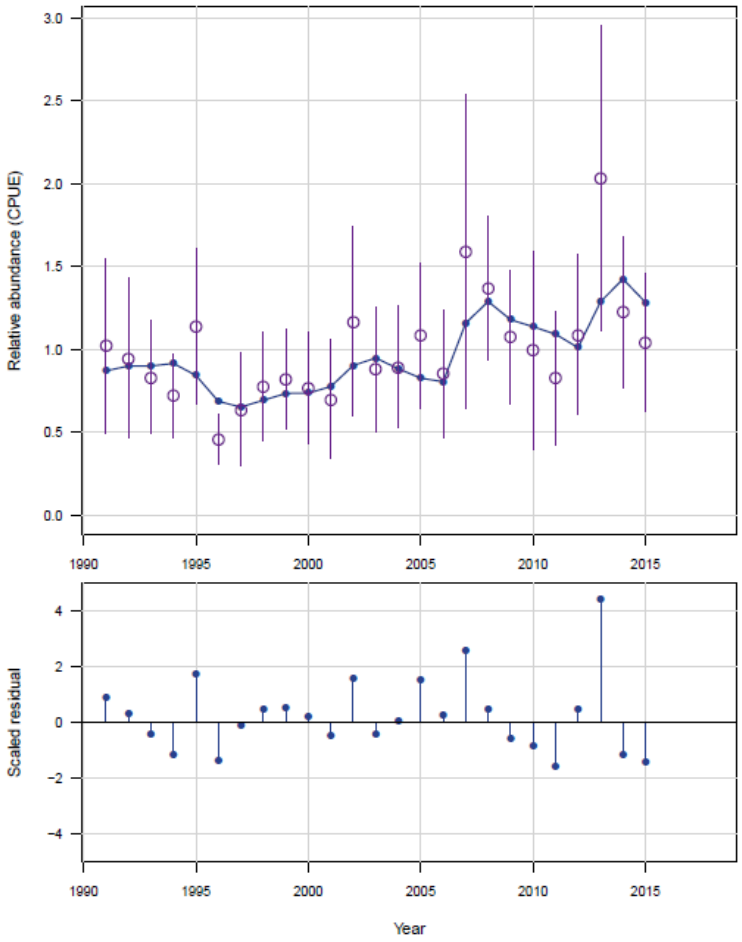

Figure 2: Time series example. Relative abundance (CPUE) time series with observed and expected values (upper panel) and scaled residuals over time (lower panel) for Atlantic cobia. Source: SEDAR, 2020a.

patterns. For example, plots of residuals versus time/fitted values/covariates/omitted variables of datasets (Fig. 2) help analysts identify correlations between poor model fit and time/fitted values/etc.

If there are strong residual patterns, there may be fishery or population dynamics (time-varying selectivity or growth) that the current model parameterization is failing to account for, particularly if the magnitude of the residual pattern is correlated with time. The 2020 Pacific sardine benchmark assessment, for example (Kuriyama et al., 2020), included a time-varying age-based selectivity form to accommodate dynamic movements and migrations leading to high year-to-year variability in the fish available to fishing fleets, leading to age compositions that can be quite variable. The tradeoff, however, is that the model may not always be able to reliably estimate the increased number of parameters compared to a model without time-varying selectivity, resulting in model misspecification. Residual analysis is just one component of model development and should always be interpreted in the context of knowledge of the fishery system.

There are many logistical challenges associated with fishery data. Thus it is unreasonable to expect perfect model fits. It is an analyst's responsibility to evaluate a range of model configurations and hypotheses to develop a model with acceptable distributions of residuals.

3.2.2. Likelihood

Integrated stock assessment models with joint likelihood functions typically incorporate many types of data to characterize biological processes (e.g., stock-recruit relationships, growth, movement). A major challenge with these integrated models, however, is the influence of the different data types on the likelihood of estimates of abundance or population trends. The different data types may provide different and sometimes conflicting information to the joint likelihood of the model.

Likelihood profiles, in which a parameter of interest (e.g., steepness, natural mortality, equilibrium recruitment) is fixed at different values and the model is re-estimated, allow the analyst to identify the relative information in each data type and get a sense of the likelihood surface surrounding MLEs. Likelihood profiles can identify data sets with conflicting information (e.g., see "Piner plots" in r4ss; Fig.

3) and evaluate model sensitivities. Values that fall within the 95% confidence interval (dashed horizontal line) are considered to be supported by the data. Data conflicts are indicated when the objective components of different data sources achieve minima at different values for a given parameter. When this occurs, the parameter estimate is sensitive to the relative weighting among data sources, and therefore careful data weighting becomes even more critical. When the profile is flat and/or the parameter is minimized at a bound (e.g., bottom panel in Fig. 3) it suggests that there is an inability to estimate the parameter from any of the data sets and that the parameter should potentially be fixed, as it was in the Pacific sardine stock assessment (Kuriyama et al., 2020).

Univariate likelihood profiles may not account for correlation among parameter estimates, and it is recommended to evaluate the estimated correlation matrix for confounded parameters with high correlations. Bivariate likelihood profiles, where two parameters are fixed across a range of values and the model is rerun for each combination of the fixed parameters, are time consuming but can be informative, especially when parameters are correlated. For vermilion snapper in the Gulf of Mexico, profiles were carried out for a combination of steepness and stock-recruit variance parameters, and contour plots (Fig. 4), where the color scale provides the negative log-likelihood value, were used to determine the relationship between the parameters. Although the final model estimates of σ_R (0.3; eventually fixed at this value in the base model) and steepness (0.71) provide the smallest negative log-likelihood value, a number of alternate pairings give approximately similar negative log-likelihood values (SEDAR, 2020b).

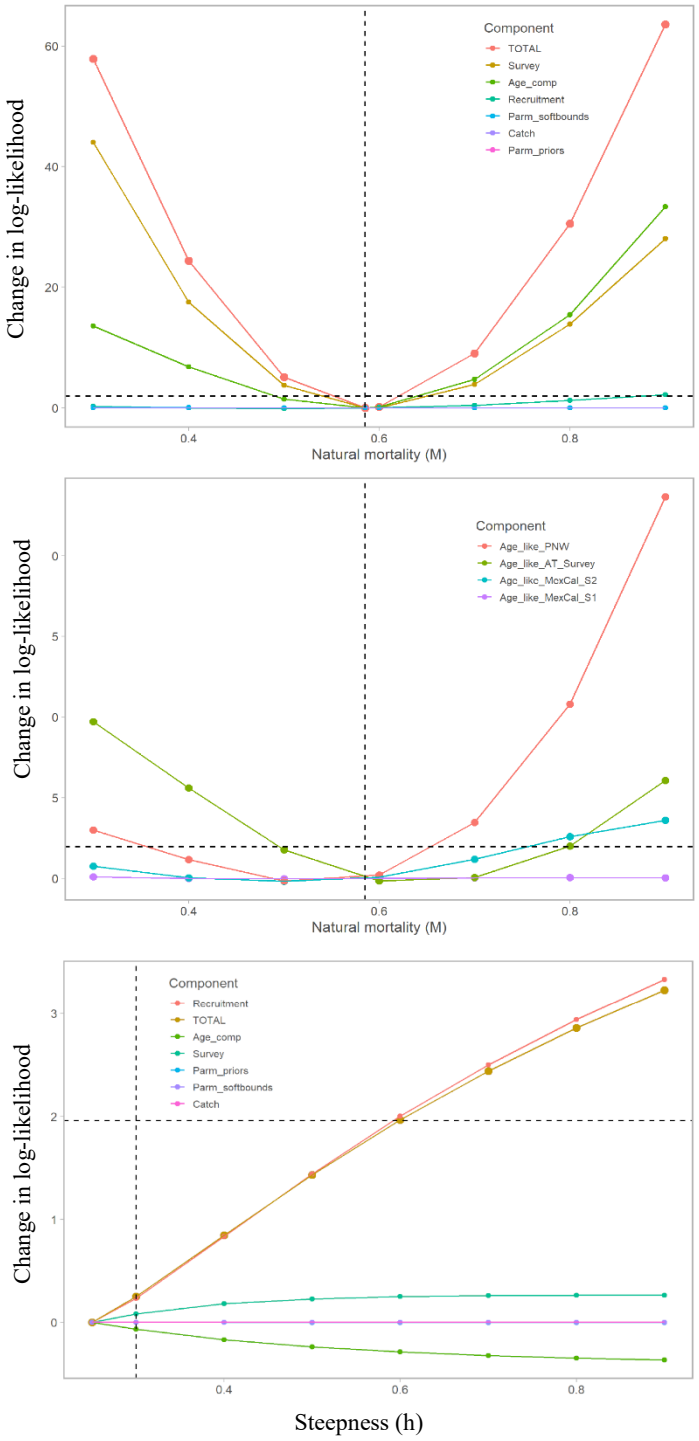

Figure 3: Likelihood profile across fixed values of natural mortality (M) (top, middle panel), and steepness (bottom panel) from the 2020 Pacific sardine stock assessment (Kuriyama et al., 2020).

When comparing likelihoods, it is important to evaluate the total likelihood and individual likelihood components associated with each data type (e.g., indices, length compositions, age compositions) and data

source (e.g., multiple fleets of a specific data type). When making these comparisons, analysts may prioritize one data type more than another based on prior knowledge. Comparison of likelihoods between models with different data weights (e.g., different effective sample sizes) should be avoided, as this rescales the likelihood and prevents direct comparison.

When data conflicts are detected, this might indicate sampling error in a data component or that the model is misspecified. If the model is misspecified, additional or alternative structure/complexity may be necessary. An analyst can therefore try adding more or less complexity to the model to minimize the total likelihood and likelihoods associated with each data source. Increasing the complexity of a model will increase the number of parameters and likely result in a decrease in likelihood values. Metrics such as Akaike Information Criterion (AIC; Akaike, 1998) are one means of avoiding overfitting and accurately

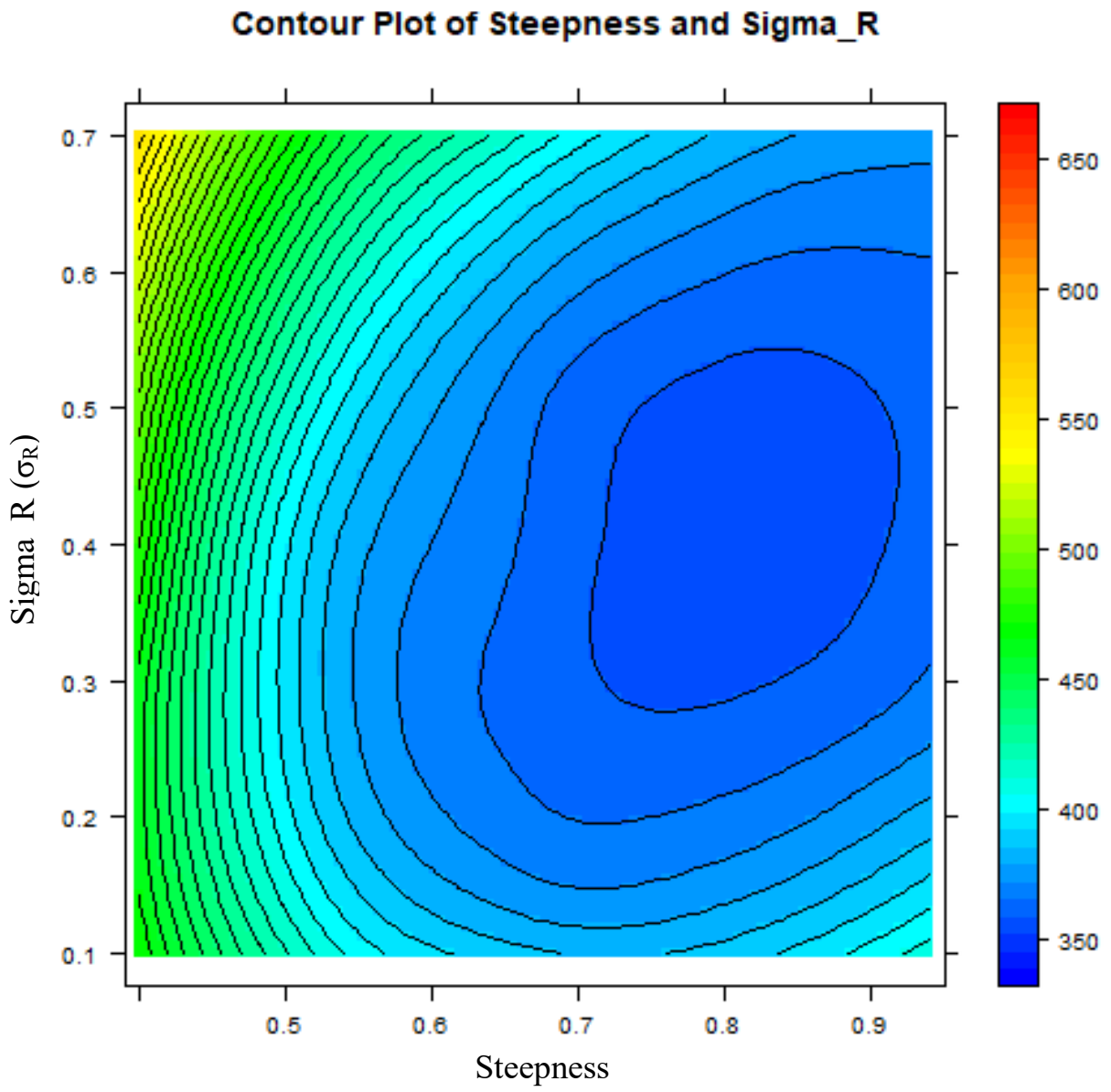

Figure 4: Contour plot of steepness and the standard deviation of recruitment variability (σ_R) for Gulf Vermilion Snapper (SEDAR, 2020b).

characterizing a population's dynamics. Making a model more complicated, however, is not always the appropriate solution, especially if the issues are due to sampling errors and not model misspecification. Often, when a model fails a diagnostic test, it is assumed to be an issue with the model, but it is also important to consider data issues in these situations. Indications of sampling error can be critical to the assessment model, but can also be among some of the most difficult to handle. Some approaches to dealing with these data issues (if identified) include, but are not limited to: dropping the data source, modifying the input data, or down weighting the data input.

3.3. Model Consistency & Sensitivity Analyses

3.3.1. Basic robustness to assumptions

Assessments contain many different inputs and structural assumptions. Various types of sensitivity analyses can be used to evaluate the response (i.e., the model's basic robustness) to changes in model input and model structure, including key assumptions related to biology, fishing mortality, population structure, and observation and process errors. This can be useful for characterizing uncertainty in the assessment or simply for a better understanding of model behavior. While many U.S. assessments are considered data-rich, there can still exist uncertainty in various processes and the differences among models considered in sensitivity analyses often show greater variability than the internally estimated variances of parameters and derived quantities in the base model (Fig. 5). Characterizing this broader level of uncertainty is one motivation for ensemble modeling.

Sensitivity analyses should evaluate the influence of alternative assumptions on model results. Sensitivity analyses are a key component of U.S. assessments used in all regions to quantify the uncertainty associated with estimating (or fixing) different parameters or alternative model parameterizations, and subsequently evaluating model results. Sensitivity analyses may include model runs with a range of assumptions. Examples of situation-dependent assumptions might include change points in time (e.g., gear switching for a fishing fleet), population structure, stock-recruitment, growth, density-dependence, and maturation. Spatial structure is another assumption that may be explored, but often this requires a large amount of data (e.g., bluefin tuna which spans the Pacific Ocean) and even in these cases not much structure may be estimable.

The first step to running a sensitivity analysis is to identify key uncertainties that can be practically evaluated, and then to provide plausible scenarios. In addition, it is important to identify the reason for each sensitivity run. Is it proposed as a possible alternative state of nature, or is it simply to better understand model behavior? Exploration and analysis of these elements of the assessment are often the focus of review panel meetings. For example, the 2021 Lingcod STAR Panel report[9] had 20 panel requests that involved adding time-varying selectivity parameters, estimating early recruitment deviations, and fixing the female natural mortality parameter.

3.3.2. Age-structured production model

One specific sensitivity analysis tool, commonly used for highly migratory species assessments primarily at the SWFSC and PIFSC, is an age-structured production model (ASPM). The ASPM diagnostic was proposed by Maunder and Piner (2015) to further evaluate model misspecification and ascertain the influence of composition data on the estimates of trends and absolute abundance. In its essence, the

[9] https://www.pcouncil.org/documents/2021/10/lingcod-stock-assessment-review-star-panel-report.pdf/

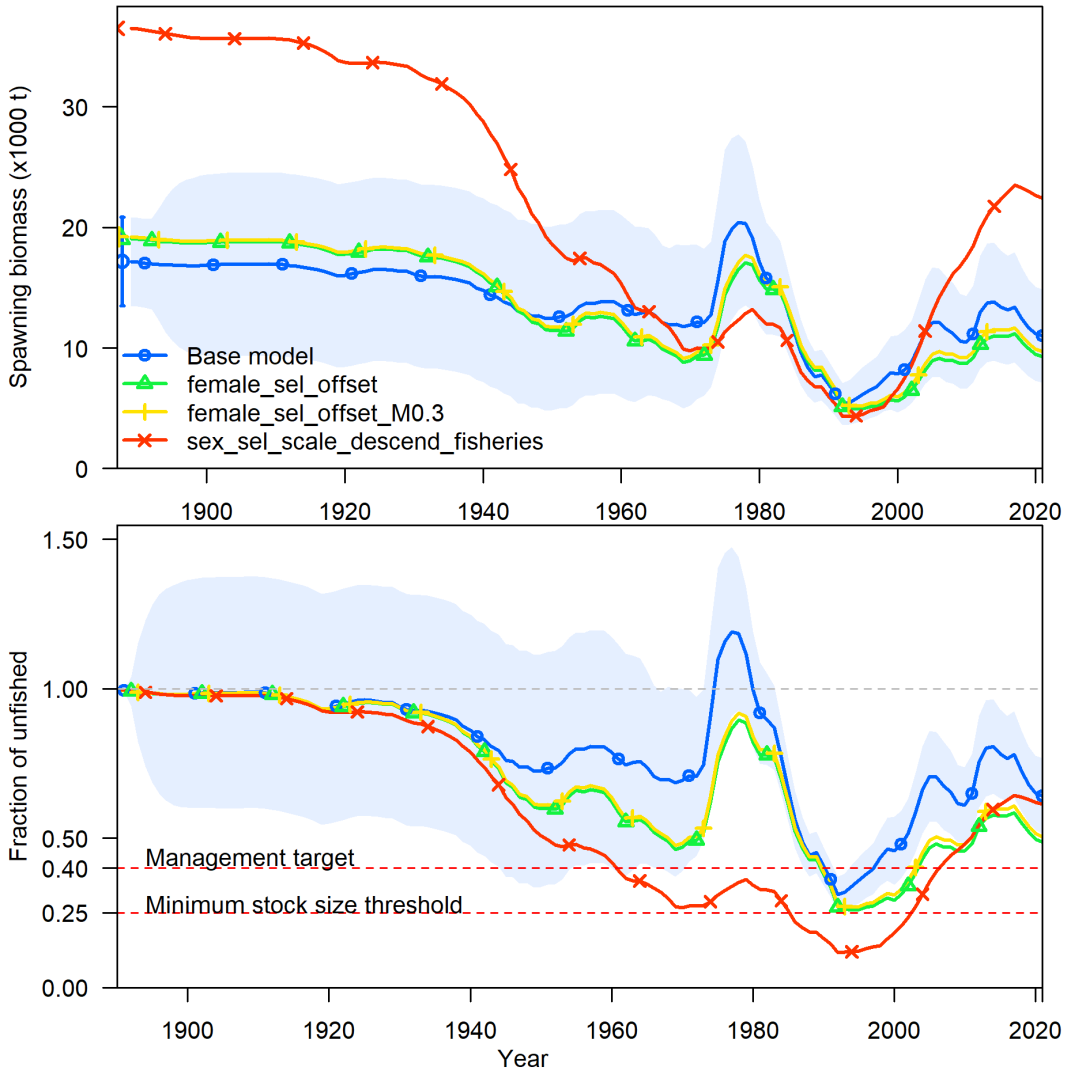

Figure 5: Example where sensitivity analyses to alternative assumptions (in this case related to parameterization of sex-specific selectivity for lingcod) shows greater variability in spawning biomass (top) and fraction unfished (bottom) than the uncertainty associated with the base model (blue shading) (from Taylor et al., 2021).

ASPM can be used for assessing whether surplus production and the observed catches alone can explain the trends in the indices of abundance.

The ASPM diagnostic is computed as follows: (i) run the fully integrated model; (ii) fix selectivity parameters at the MLE from the fully integrated model, (iii) turn off the estimation of all parameters except the scaling parameters and the parameters representing the initial conditions (a parameter for the equilibrium recruitment and a parameter for the equilibrium fishing mortality), set the recruitment deviates to zero (early recruitment and model period recruitments); (iv) fit the model to the indices of abundance only; and (v) compare the estimated trajectory to the one obtained in the base case.

If the ASPM is able to fit the indices of abundance that have good contrast (i.e., those that have declining and/or increasing trends) then Maunder and Piner (2015) suggest that this is evidence of a production function's existence, and the indices likely provide information about absolute abundance. They refer to this situation as "the catch explains the indices well." Subsequently, an ASPM with recruitment

deviations estimated (ASPMrec) can be applied to evaluate whether temporal variability in recruitment can be estimated without using age- or size-composition data directly.

If the ASPM does not fit the indices well, that is an indication that the catch alone cannot explain the index trends. This can have several causes, including that the stock is recruitment-driven, or the indices of relative abundance are not proportional to abundance. Checking whether the stock is recruitment-driven involves fitting the ASPMrec. If the ASPMrec cannot capture the population trajectory estimated in the integrated model, it can be concluded that the information about scale in the integrated model is not coming from the CPUE data and the catches, but rather from the composition data. Composition data can often provide the best information about recruitment and selectivity, but their influence on the estimation of absolute abundance needs to be taken with caution.

An example of the use of the ASPM diagnostic is in the stock assessment for the North Atlantic shortfin mako shark (Courtney et al., 2017). Pelagic longline operations catch the vast majority of shortfin mako shark, but due to strong spatial structuring of size classes, the selectivity pattern differs among the fishing fleets operating in the different regions. The population dynamics of shortfin mako reveal an unusual combination of slow somatic growth, very late maturation, and steep dome-shaped selectivity. The shortfin mako shark example represents a length-based age- and sex-structured multi-fleet model fit to six standardized CPUE indices. Fisheries-dependent length-composition data are assumed to be representative of the different selectivity patterns for the six major surface longline fishing fleets.

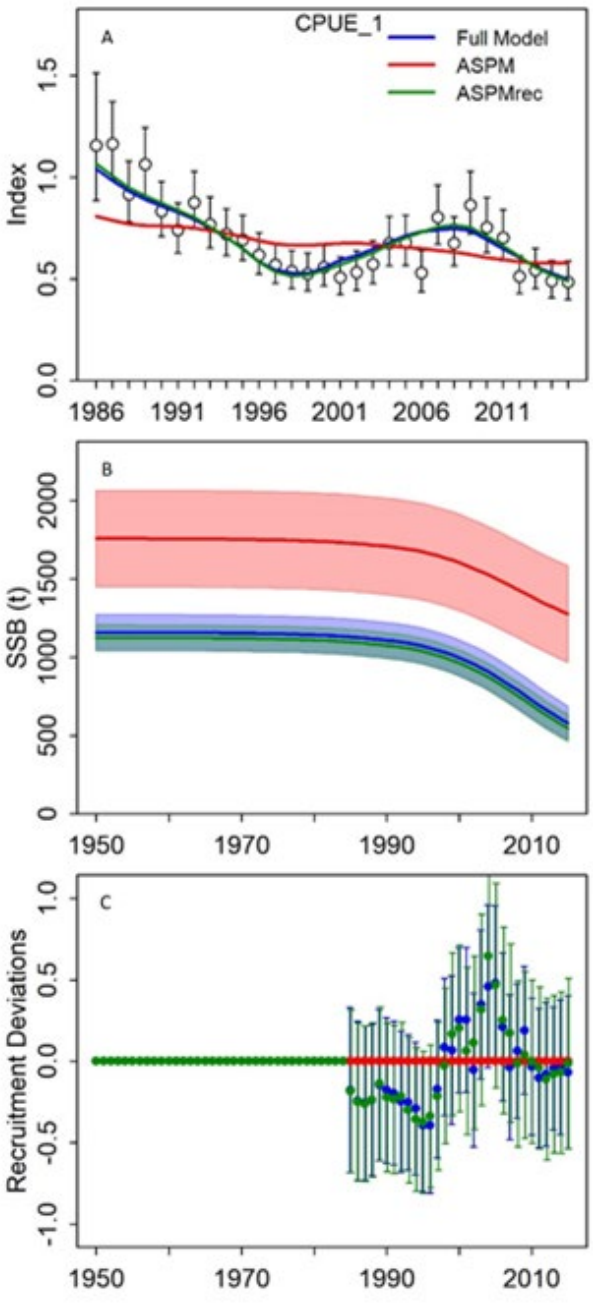

Figure 6: Comparison between the fully integrated base-case and the deterministic Age-Structured-Production Model (ASPM) results for North Atlantic shortfin mako (Courtney et al., 2017).

The CPUE trend estimated by the ASPM is very different from those estimated in the fully integrated assessment model (Fig. 6, top panel). The ASPM diagnostic showed a consistent declining trend over time. The fit to the same index in the fully integrated shortfin mako shark and ASPMrec models was almost identical and had a more oscillatory pattern (Fig. 6, top panel). The ASPM can estimate the "correct" scale of the biomass only when recruitments are allowed to vary (Fig. 6, middle panel). Results

from the ASPMrec indicated that the CPUE data and the catches contained information on temporal variability on recruitment (Fig. 6, bottom panel).

Any stock assessment must find the relationship between fishing and changes in abundance. The ASPM is straightforward to produce in software packages like SS3 (see Carvalho et al., 2021 and the associated {ss3diags} R package[10]), and its implementation should be encouraged while building a complicated assessment model.

To date the ASPM diagnostic is widely used for highly migratory species assessments at the SWFSC and PIFSC, but has only rarely been applied to groundfish or reef fish stocks. The application to Pacific Hake by Stewart and Forrest (2011), Pacific ocean perch[11], and blueline tilefish in the South Atlantic in SEDAR 50 (SEDAR, 2017) provide a few examples.

3.3.3. Leave-one-out analysis

A leave-one-out analysis (sometimes referred to as a "jack-knife" analysis) is used to determine if any single data source (or data point) is having undue influence on the model estimates and causing tension with other data in terms of estimating parameters. Leave-one-out analysis involves removing individual data sets one at a time and refitting the model to the remaining data. This analysis can be used to evaluate the stability of the assessment (i.e., whether the addition or the change of a data source leads to different model estimates). Leave-one-out analysis can also be conducted to identify data points with high leverage and to evaluate the predictive capability of the assessment model (Brooks and Deroba, 2015). If removing a data point leads to dramatically different results, the data point should be re-evaluated with the data providers to ensure it reflects the best available data.

Leave-one-out analyses of indices of abundance are often conducted to determine if any single index is driving model estimates of derived quantities (e.g., spawning biomass, recruitment). Each index of abundance is removed one at a time, and results can identify indices that may be giving conflicting abundance trend signals compared to the remaining indices. Additionally, groups of indices may be removed simultaneously (e.g., all recreational CPUE indices, all commercial CPUE indices, all fishery-dependent CPUE indices, and all fishery-independent surveys). A full leave-one-out analysis would include removing all information associated with a survey (index of abundance, including associated age or size composition data) one survey at a time and refitting the model.

Leave-one-out analyses can be thought of as sensitivities and requires the following steps: (i) run the integrated assessment model; (ii) remove one data source at a time and refit (i); and (iii) repeat (ii) for each data source. Within SS3, this analysis can be conducted by multiplying the likelihoods for a given fleet times zero such that they do not inform parameter estimates. For surplus production models that are fit to more than one abundance index, individual indices are removed one at a time. It may also be possible to conduct a 'leave-one-in' analysis, fitting the model to only one index at a time. Note that if a data source is removed, the analyst must also turn off estimation of any parameters that depend entirely on that data source (e.g., catchability associated with an index). Trends in year-specific estimates of abundance, biomass, recruitment, and mortality can then be compared to determine if any one data source is greatly impacting the model. The effects of removing composition data on growth can also be examined. The expected outcome is that no one data source will be driving the assessment results, and that the results will generally be in agreement when each data source is removed. If removing a dataset

[10] https://github.com/PIFSCstockassessments/ss3diags
[11] https://www.pcouncil.org/documents/2017/06/pacific-ocean-perch-star-panel-report-26-30-june-2017.pdf/

leads to dramatically different results, the dataset should be reexamined to determine if the sampling procedures are consistent and appropriate (e.g., an index may only be sampling a subunit of the stock and resulting abundance signals may only reflect a local sub-population and not the trend in the entire stock).

Regional Practices and Examples

Leave-one-out analyses are handled differently among regions. In the Southeast, leave-one-out analyses are standard sensitivity runs conducted for Gulf of Mexico reef fish assessments. They often focus on removing one index of abundance at a time, or a group of indices. Full leave-one-out analyses are usually not conducted because the other data sets are considered fundamentally necessary to stabilize the assessment. Generally, a leave-one-out analysis is not used as a pass/fail criterion for a base assessment model, and no model adjustments are made based on the results. It is primarily used to provide scientists and managers with an understanding of how sensitive the model outcomes are to the indices of abundance that are included in the model, given the high volume of indices incorporated into Southeast assessments (range: 4 [Gulf yellowedge grouper, Gulf tilefish] to 18 [Gulf red snapper]). For example, the removal of the video index for the Gulf vermilion snapper assessment had a noticeable impact on both the estimate of spawning output and recruitment in the last few years of the assessment (Fig. 7). While this was discussed in detail during the assessment process and by the Gulf SSC, ultimately the base SEDAR67 model (including all indices of abundance) was used to set catch advice (SEDAR, 2020b).

Another form of leave-one-out analysis applied in the SEFSC focuses on recreational removals. Time series of recreational landings and discards commonly display "spikes," in which a single value (observed) is ~4-5 times larger than those in surrounding years. The veracity of these spikes is routinely questioned, as is their effects on assessment output and projections. Thus, these effects are explored through sensitivity analyses where the spike is left out and replaced by a local average.

At the NEFSC, leave-one-out analyses are conducted in a similar manner as the SEFSC and are generally limited to considering the impact of removing abundance indices from the assessment. While not mandated in the Terms of Reference, leave-one-out analyses are often completed in assessments that incorporate several indices of abundance in order to evaluate the sensitivity of model estimates to the included indices. However, leave-one-out analyses have not been used to inform dataset weightings or which indices are included in the assessment model.

At the NWFSC and SWFSC, leave-one-out analyses vary among assessments where some assessments will emphasize investigating the removal of a single data type (e.g., the index) and more data-rich assessments will perform a full leave-one-out analysis. Few assessments to date have explored the removal of individual data points within a data set. When used for management, surplus production models often perform both leave-one-out and leave-one-in analyses to better understand conflicts in trend information (Fig. 8; Dick and MacCall, 2014). Highly migratory species assessments use jackknifing in the development of abundance indices but not in the formal assessment. Coastal pelagic species do not include jackknifing nor leave-one-out analyses.

The PIFSC does not typically use jackknife-type analyses for characterizing uncertainty in stock assessments. This is primarily due to the fact that there are few relative abundance indices available for most stock assessments. However, many PIFSC assessments include sensitivity analyses where one or more abundance index or size composition likelihood components are excluded from the model fitting process. This sensitivity information is provided to characterize the effects of removing one or more data sources on model diagnostics and results.

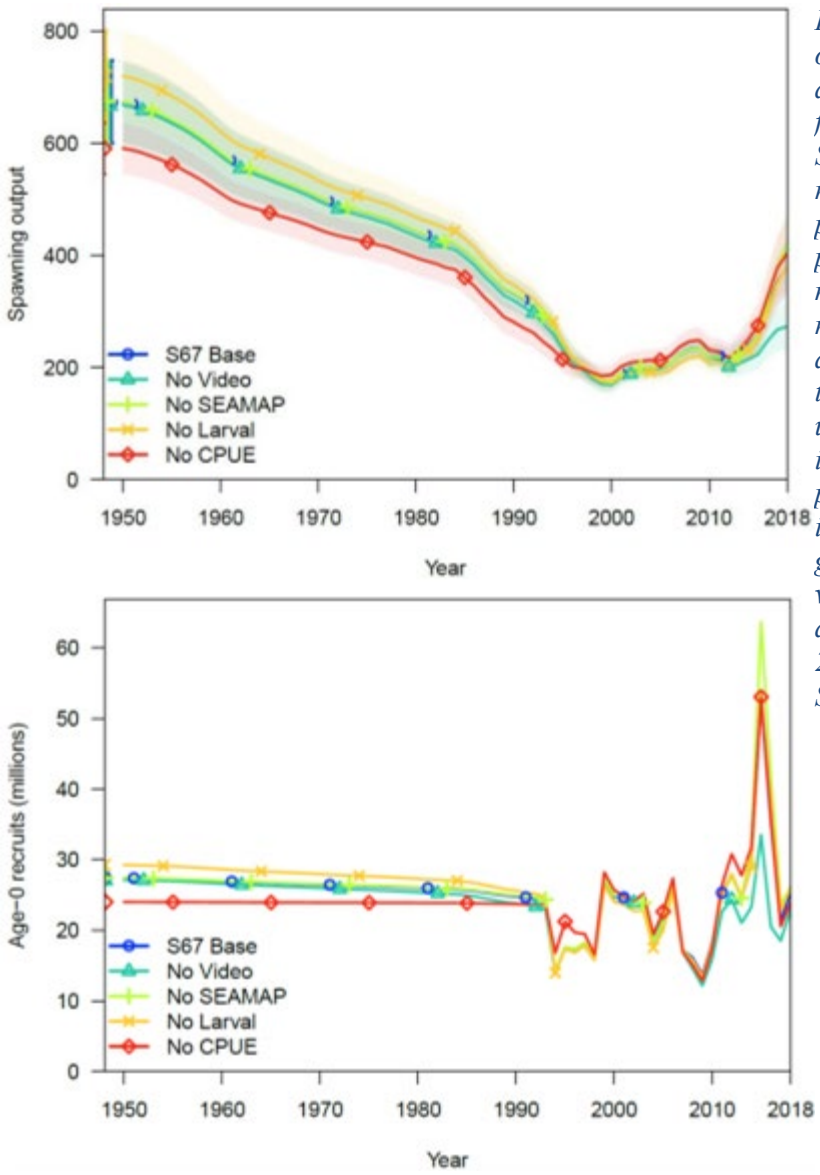

Figure 7: Results of a leave-one-out analysis with the fishery-dependent and independent indices for Gulf vermilion snapper. Spawning stock biomass and recruitment (million fish; bottom panel) are shown. The analysis was performed by running the base model with one of the indices removed (or all of the fishery-dependent CPUE indices) in order to determine if any given index had undue influence on model results or indicated widely differing trends in population trajectories. The results indicate most of the indices are generally in agreement, but the video index appears to be a strong driver in estimating the extreme 2015 recruitment event. Source: SEDAR, 2020b.

Staff at the PIFSC are currently investigating the use of k-fold cross validation to estimate relative model weights for ensemble modeling, where we note that the k-fold cross validation procedure is similar to the delete-m jackknife procedure. Here the idea is to quantify the individual model fits to the hold-out data subset using a specific distribution that can be described by a log-likelihood function. The sum of the log-likelihood fits over the k-fold data subsets can provide a measure of predictive accuracy for each model in the ensemble, and these measures can be used to compute model weights (Hauenstein et al., 2018) that are similar to the AIC weights for multimodel inference described in Burnham and Anderson (2002). This approach is very general and can be applied with any model fitting algorithm (e.g., maximum likelihood, Bayes, random effects, machine learning, etc.) provided there is an appropriate log-likelihood for the predicted data.

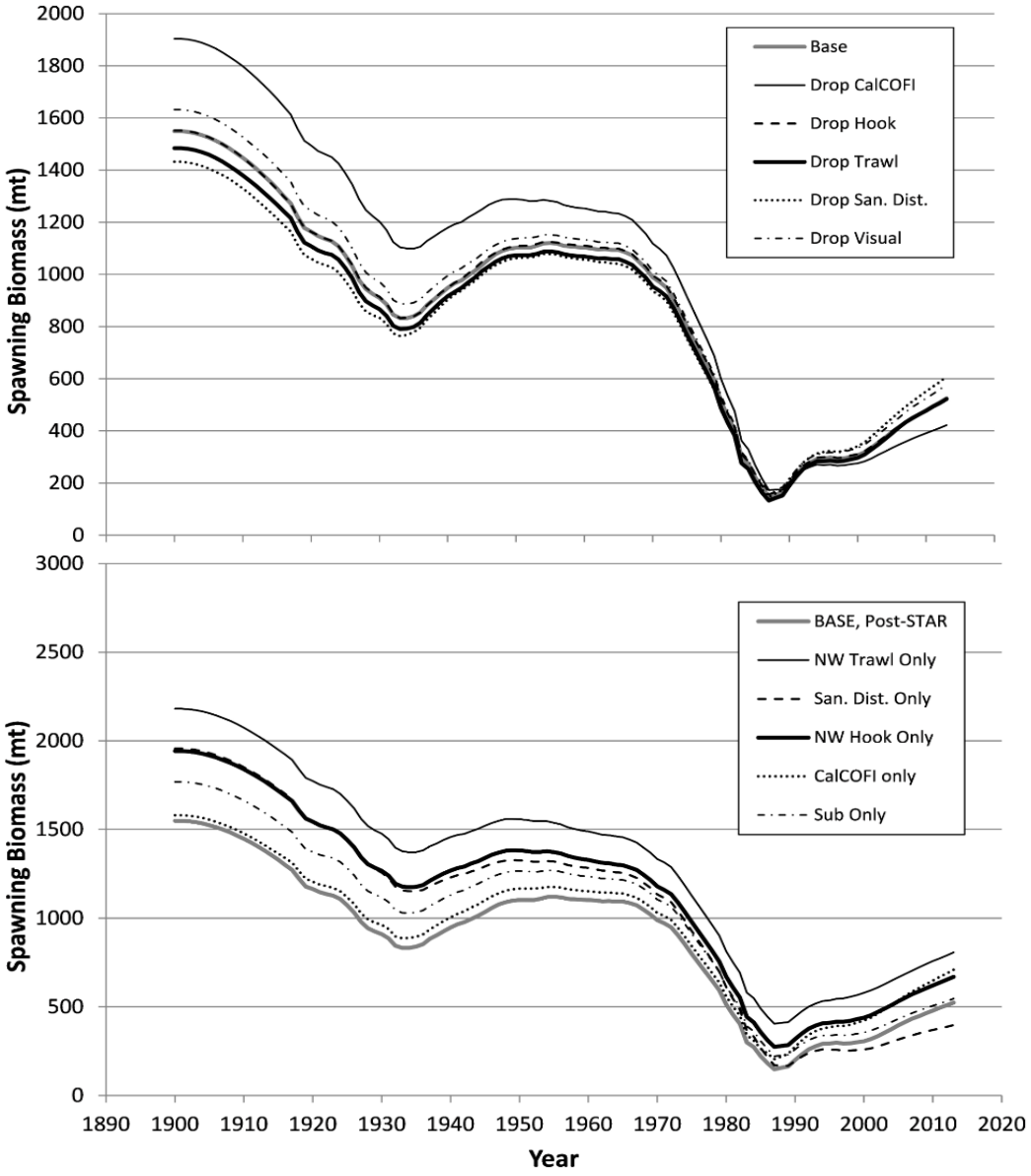

Figure 8: Median spawning biomass for cowcod surplus production model fit by excluding individual indices ('leave-one-out', upper panel), or including individual indices ('leave-one-in', lower panel. Source: Dick and MacCall, 2014.

3.3.4. Retrospective Analysis

A retrospective analysis is used to determine whether there is a misspecification in the estimation model (ICES, 2020). It assesses the impact of the most recent years of data on model estimates and determines whether estimated quantities are consistently over- or underestimated.

A retrospective pattern is a systematic inconsistency among a series of model estimates such as recruitment, population size, spawning stock biomass, or fishing mortality, based on increasing years of data (Mohn, 1999). Retrospective patterns often arise due to a temporal change in life-history characteristics (e.g., natural mortality, growth), selectivity, or in the accuracy of input data (e.g., fishery

removals) that are misspecified in the model (Legault, 2009; Deroba, 2014; Hurtado-Ferro et al., 2015). A within-model retrospective analysis is useful for determining an internal inconsistency in the data because the only change between model runs is the number of years of data (Legault, 2009).

The following steps are necessary for computing a retrospective analysis: (i) run the integrated assessment model; (ii) remove the terminal year of data, including catches, and rerun the model to the reduced terminal year; (iii) repeat step 2 for a total of at least 5-7 years. Trends in year-specific estimates of recruitment, abundance, biomass, and mortality can then be compared to results from the base model with all years of data to determine if they are consistently over- or underestimated. Note that particular care should be taken if the assessment model has time blocks on selectivity or natural mortality when conducting retrospective analyses.

The severity of a retrospective pattern can be evaluated using Mohn's rho, which is defined as the average of the relative differences between estimates from a model using a truncated time series and estimates from one based on full time series (Mohn, 1999; Hurtado-Ferro et al., 2015). A positive Mohn's rho value indicates that the estimated quantity is consistently overestimated as years of data are removed. If a retrospective pattern (consistent over- or underestimation of biomass, abundance, and/or mortality across model runs) exists, a retrospective adjustment to model estimates can be made, and it is typically based on Mohn's rho. For example, $\hat{\theta} = \theta \frac{1}{1+\rho}$, where $\hat{\theta}$ is the adjusted value, θ is the unadjusted value, and ρ is Mohn's rho (see Legault, 2009). However, ideally before any adjustments to model estimates are made, the model parameterization should be re-evaluated to determine if adjustments to the model parametrization could address the retrospective pattern.

Regional Practices and Examples

Retrospective patterns are handled differently among regions. At the NEFSC, if a "major" retrospective pattern exists, defined as an adjustment that shifts the terminal year fishing mortality or spawning biomass outside the 90% confidence bounds of the original estimates, a retrospective adjustment is applied to model results for determining stock status as well as for projections. In some extreme cases, however, the retrospective pattern can result in the rejection of the age-structured assessment model, as was the case for the Yellowtail flounder (Legault et al., 2014) (Fig. 9) and Witch flounder (NEFSC, 2017) assessments.

For AFSC assessments, all groundfish and crab models are required to produce a 10 year retrospective analysis on female spawning biomass (or mature male biomass for crabs). Guidance explicitly states that Mohn's rho in isolation is not a cause for rejection of a model or an adjustment, but rather to be used in evaluating alternative models (Hanselman et al., 2013). The presence of a large positive Mohn's rho has been used to justify recommending more precautionary catch advice in some cases or to choose a different model alternative. In some cases, evaluating the sequence of tuning components of the retrospective analyses can reveal the cause (e.g., Lowe et al., 2018).

Retrospective analysis is also required for all SWFSC and NWFSC groundfish and coastal pelagic species assessments. Highly migratory species assessments evaluate retrospective patterns, although this is not a formal requirement for problematic retrospective patterns. Coastal pelagic species assessments generally have short modeling periods (10-15 years) and require strong assumptions on biological parameters. This results in a negligible retrospective pattern. However, the results of the analysis are not formally incorporated into management decisions. Relatively few west coast groundfish assessments show strong retrospective patterns.

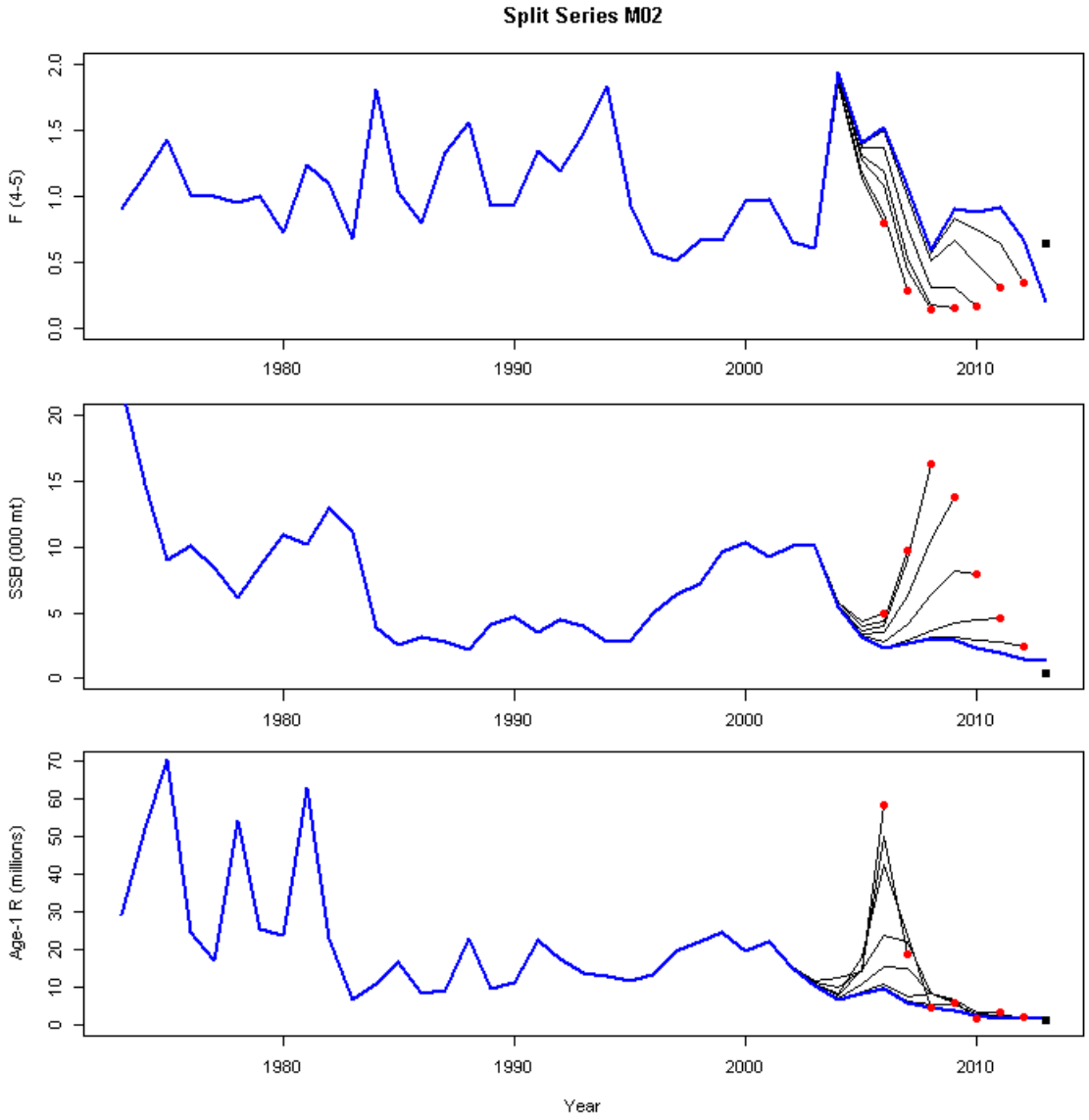

Figure 9: Summary plots showing strong retrospective patterns in fishing mortality (F, top panel), spawning biomass (SSB, middle panel), and recruits (age-1 R, bottom panel) from the Yellowtail flounder assessment (Legault et al., 2014) as successive years of data are excluded from the assessment.

Retrospective analysis is commonly used as a model diagnostic for both integrated assessment models and Bayesian biomass dynamics models at the PIFSC. A typical analysis uses a time period of 5 years of retrospective peels to evaluate whether there is a consistent retrospective pattern for providing management advice. Retrospective patterns are evaluated for estimates of spawning biomass and fishing intensity, or fishing mortality, using Mohn's rho (Mohn, 1999) or similar diagnostic (Hurtado-Ferro et al., 2015) as a measure of the strength of the retrospective pattern. The direction and strength of the retrospective patterns in spawning potential and fishing intensity are reported as model diagnostics and used for management advice.

For example, in the 2021 stock assessment of Pacific blue marlin (ISC, 2021), there were two equally-weighted SS3 models in the model ensemble classified as best scientific information available. Then Mohn's rho was calculated for the biomass and fishing mortality peels, and the severity of the retrospective pattern was based on the range provided by Hurtado-Ferro et al. (2015), with values higher than 0.20 and lower than -0.15 used as an indication for problematic retrospective patterns. Both models exhibited a relatively strong retrospective pattern in recent years, with Mohn's rho values for spawning biomass and fishing mortality of about ρ(spawning biomass) ≈ 0.3 and ρ(fishing mortality) ≈ -0.3, respectively. In this case, the conservation advice included a statement for managers to consider when making management decisions that there is an apparent tendency to overestimate spawning biomass and underestimate fishing mortality, in part due to an unusual decline in Japanese longline CPUE in recent years.

At the SEFSC, retrospective analysis is routinely used to assess the consistency of terminal year model estimates. Generally, this analysis is not used as a pass/fail criterion, and results for 5- or 10-year retrospective analyses are presented to managers to consider any additional uncertainties when making decisions. If the resulting estimates of derived quantities such as spawning biomass or recruitment differ significantly, particularly if there is serial over- or underestimation of any important quantities, the model may have some unidentified process error, which requires reassessing model assumptions.

Retrospective analysis is also a commonly used diagnostic tool for European assessments within the International Council for the Exploration of the Sea (ICES) community. A 2020 ICES report provides recommendations on retrospective patterns which may also prove useful in U.S. stock assessment context, and we point readers to this report for a more in-depth discussion on retrospective analysis (ICES, 2020).

4. CONCLUSIONS & FUTURE DIRECTIONS

The goal of this document is to provide a description of the diagnostics that are most top-of-mind for stock assessment scientists from around the country. This document is by no means comprehensive, and we suggest readers also consult Carvalho et al. (2021) and reports from the Center for the Advancement of Population Assessment Methodology[12].

Stock assessment methodology is developing rapidly, and these developments will be necessary given a future with a changing climate and dynamic environmental conditions. Many of the diagnostics here will likely continue to be relevant, even with more advanced methods and inclusion of additional data sources. The next generation of U.S. stock assessment models are currently in development as part of the NOAA Fisheries Integrated Modeling System working group. The group is focusing on more explicit exploration and integration of socioeconomic factors and environmental drivers in stock assessments and transitioning to a modular and extensible software that can further leverage high performance and cloud computing. Despite future advances in software and methodology, stock assessment scientists will continue to rely on a strong understanding of the data and a suite of diagnostics to identify model sensitivities to ensure that fisheries management is informed by the best scientific information available.

5. LITERATURE CITED

Akaike H. 1998. Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. 199 p. Springer, Berlin, Germany.

[12] See Diagnostics Workshop, capamresearch.org

Brooks, E. N., and J. J. Deroba. 2015. When "data" are not data: the pitfalls of post hoc analyses that use stock assessment model output. Canadian Journal of Fisheries and Aquatic Sciences. 72(4): 634-641.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach, 2nd ed. Springer, New York, NY.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., and Betancourt, M. 2017. Stan: A Probabilistic Programming Language. Journal of Statistical Software. 76(1): 1-32. https://doi.org/10.18637/jss.v076.i01

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

Courtney, D., E. Cortes, and X. Zhang. 2017. Stock Synthesis (SS3) model runs conducted for North Atlantic shortfin mako. Collect. Vol. Sci. Pap. ICCAT. 74(4):1759–1821.

Deroba, J. J. 2014. Evaluating the consequences of adjusting fish stock assessment estimates of biomass for retrospective patterns using Mohn's rho. North American Journal of Fisheries Management. 34:380–390.

Dick, E. J,. and A. MacCall. 2014. Status and productivity of cowcod, Sebastes levis, in the Southern California Bight, 2013. Pacific Fishery Management Council. Available from http://www.pcouncil.org/groundfish/stock-assessments/

Edwards, A. M., A. M. Berger, C. J. Grandin, and K. F. Johnson. 2022. Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2022. Prepared by the Joint Technical Committee of the U.S. and Canada Pacific Hake/Whiting Agreement, National Marine Fisheries Service and Fisheries and Oceans Canada. 238 p.

Gabry J, Veen D (2022). _shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models_. R package version 2.6.0, https://mc-stan.org/users/interfaces/shinystan.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian Data Analysis, 3rd ed. Chapman and Hall Book, CRC Press, Boca Raton, LA. https://doi.org/10.1201/b16018

Haddon, M. 2011. Modelling and quantitative methods in fisheries, 2nd ed., 465 p. Chapman & Hall Book, CRC Press, Boca Raton, LA.

Hanselman, D. H., B. Clark, and M. Sigler. 2013. Report of the groundfish plan team retrospective investigations group. Available at http://www.afsc.noaa.gov/REFM/stocks/Plan_Team/2013/Sept/Retrospectives_2013_final3.pdf

Hauenstein, S., S. N. Wood, and C. F. Dormann. 2018. Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation. Communications in Statistics-Simulation and Computation, 47(5):1382-1396.

Hurtado-Ferro, F., C. S. Szuwalski, J. L. Valero, S. C. Anderson, C. J. Cunningham, K. F. Johnson, R. Lican-deo, C. R. McGilliard, C. C. Monnahan, M. K. Muradian, K. Ono, K. A. Vert-Pre, A. R. Whitten, and A. E. Punt. 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science. 72:99–110.

ICES. 2020. Workshop on Catch Forecast from Biased Assessments (WKFORBIAS; outputs from 2019 meeting). ICES Scientific Reports. 2:28. 38 p. http://doi.org/10.17895/ices.pub.5997

ISC. 2021. Stock Assessment Report for Pacific Blue Marlin (*Makaira nigricans*) Through 2019. 21st Meeting of the International Scientific Committee for Tuna and Tuna-like Species in the North Pacific. Held Virtually, July 12-20, 2021. ISC/21/ANNEX/10. Available at https://isc.fra.go.jp/pdf/ISC21/ISC21_ANNEX10_Stock_Assessment_for_Pacific_Blue_Marlin.pdf

Kuriyama, P. T., J. Z. Zwolinski, K. T. Hill, and P. R. Crone. 2020. Assessment of the Pacific sardine resource in 2020 for U.S. management in 2020-2021. NOAA Tech. Memo. NMFS-SWFSC-628.

Langseth, B., J. Syslo, A. Yau, M. Kapur, and J. Brodziak. 2018. Stock Assessment for the Main Hawaiian Islands Deep 7 Bottomfish Complex in 2018, with Catch Projections Through 2022. NOAA Tech. Memo. NMFS-PIFSC-69, 218 p. Available at https://repository.library.noaa.gov/view/noaa/17252

Legault, C. M. 2009. Report of the retrospective working group, January 14–16, 2008, Woods Hole, Massachusetts. National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Northeast Fisheries Science Center Reference Document 09-01, Woods Hole, Massachusetts. Available at https://archive.nefmc.org/tech/council_mtg_docs/Sept%202009/Herring/Doc%209_Retro%20Working%20Group%20Report.pdf

Legault, C. M., L. Alade, W. E. Gross, and H. H. Stone. 2014. Stock assessment of Georges Bank yellowtail flounder for 2014. TRAC Ref Doc. 2014/01. 214 p.

Lowe, S., J. Ianelli, and W. Palsson. 2018. Stock assessment of Aleutian Islands Atka mackerel. *In* Stock Assessment and Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions. North Pacific Fisheries Management Council, Anchorage, AK.

Lynch, P. D., R. D. Methot, and J. S. Link (eds.). 2018. Implementing a Next Generation Stock Assessment Enterprise. An Update to the NOAA Fisheries Stock Assessment Improvement Plan. NOAA Tech. Memo. NMFS-F/SPO-183, 127 p. https://doi.org/10.7755/TMSPO.183

Maunder, M. N., and K. R. Piner. 2015. Contemporary fisheries stock assessment: many issues stillremain. ICES J. Mar. Sci. 72:7–18.

Methot, Jr, R. D., C. R. Wetzel, I. G. Taylor, K. L. Doering, and K. F. Johnson. 2022. Stock Synthesis User Manual. NOAA Fisheries, Seattle WA. Available at https://nmfs-stock-synthesis.github.io/doc/SS330_User_Manual.html.

Mohn, R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES Journal of Marine Science. 56:473–488.

Monnahan, C. C., and K. Kristensen, 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.

Northeast Fisheries Science Center. 2017. 62nd Northeast Regional Stock Assessment Workshop (62nd SAW) Assessment Report. 17-03; 822 p. https://doi.org/10.7289/V5/RD-NEFSC-17-03

Quinn, T. J., and R. B. Deriso. 1999. Quantitative fish dynamics, 560 p. Oxford University Press, Inc., New York, NY.

SEDAR 2015. SEDAR 42 – Gulf of Mexico Red Grouper Stock Assessment Report. Available at https://sedarweb.org/documents/sedar-42-final-stock-assessment-report-gulf-of-mexico-red-grouper/

SEDAR. 2017. SEDAR 50 – Atlantic Blueline Tilefish Stock Assessment Report. Available at https://sedarweb.org/documents/sedar-50-stock-assessment-report-atlantic-blueline-tilefish/

SEDAR. 2020a. SEDAR 58 – Atlantic Cobia Stock Assessment Report. SEDAR, North Charleston SC. 500 p. Available at http://sedarweb.org/sedar-58

SEDAR. 2020b. SEDAR 67 – Gulf of Mexico Vermilion Snapper Stock Assessment Report. Available at http://sedarweb.org/docs/sar/S67_Final_SAR_v2.pdf.

Stan Development Team. 2017. Stan modeling language users guide and reference manual, version 2.17.0. Current version available at https://mc-stan.org/users/documentation/

Stewart, I. J., and R. E. Forrest. 2011. Status of the Pacific Hake (Whiting) stock in U.S. and Canadian Waters in 2011. Joint U.S. and Canadian Hake Technical Working Group. Available at https://www.pcouncil.org/documents/2011/04/status-of-the-pacific-hake-whiting-stock-in-u-s-and-canadian-waters-in-2011-march-2011.pdf

Stock, B. C., and T. J. Miller. 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. https://doi.org/10.1016/j.fishres.2021.105967

Subbey, S. 2018. Parameter estimation in stock assessment modelling: caveats with gradient-based algorithms. ICES J. Mar. Sci. 75:1553-1559. https://doi.org/10.1093/icesjms/fsy044

Taylor, I. G., K. F. Johnson, B. J. Langseth, A. Stephens, L. S. Lam, M. H. Monk, A. D. Whitman, and M. A. Haltuch. 2021. Status of lingcod (*Ophiodon elongatus*) along the northern U.S. west coast in 2021. Pacific Fisheries Management Council, Portland, Oregon. 254 p.

Wald, A. and, J. Wolfowitz. 1940. On a test whether two samples are from the same population. Ann. Math. Stat. 11:147-162. http://www.jstor.org/stable/2235872