ICES Journal of Marine Science

ICES Journal of Marine Science (2015), 72(1), 19–30. doi:10.1093/icesjms/fst237

Original Article

Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods

J.J. Deroba[1]*, D.S. Butterworth[2], R.D. Methot, Jr[3], J.A.A. De Oliveira[4], C. Fernandez[5], A. Nielsen[6], S.X. Cadrin[7], M. Dickey-Collas[5,8], C.M. Legault[1], J. Ianelli[9], J.L. Valero[10], C.L. Needle[11], J.M. O'Malley[12], Y-J. Chang[13], G.G. Thompson[9], C. Canales[14], D.P. Swain[15], D.C.M. Miller[8], N.T. Hintzen[16], M. Bertignac[17], L. Ibaibarriaga[18], A. Silva[19], A. Murta[19†], L.T. Kell[20], C.L. de Moor[2], A.M. Parma[21], C.M. Dichmont[22], V.R. Restrepo[23], Y. Ye[24], E. Jardim[25], P.D. Spencer[9], D.H. Hanselman[26], J. Blaylock[27], M. Mood[27], and P.-J. F. Hulson[26]

[1]*NOAA NMFS, 166 Water Street, Woods Hole, MA, USA*
[2]*Marine Resource Assessment and Management Group (MARAM), Department of Mathematics and Applied Mathematics, University of Cape Town, University Private Bag, Rondebosch 7701, South Africa*
[3]*NOAA NMFS, 2725 Montlake Blvd. E, Seattle, WA, USA*
[4]*Centre for Environment, Fisheries and Aquaculture Science (Cefas), Lowestoft Laboratory, Pakefield Road, Lowestoft, Suffolk NR33 OHT, UK*
[5]*ICES, H.C. Andersens Boulevard 44-46, DK Copenhagen V, Denmark*
[6]*National Institute of Aquatic Resources, Technical University of Denmark, Charlottenlund Castle, 2920 Charlottenlund, Denmark*
[7]*University of Massachusetts, School for Marine Science and Technology, 200 Mill Road, Suite 325, Fairhaven, MA, USA*
[8]*Wageningen Institute of Marine Resources and Ecosystem Studies (IMARES), Haringkade 1, 1976 Ijmuiden, Netherlands*
[9]*NOAA NMFS, 7600 Sand Point Way NE, Seattle, WA, USA*
[10]*Center for the Advancement of Population Assessment Methodology (CAPAM), 8901 La Jolla Shores Drive, La Jolla, CA, USA*
[11]*Marine Scotland – Science, The Marine Laboratory, PO Box 101, 375 Victoria Road, Aberdeen AB11 9DB, UK*
[12]*NOAA NMFS, 1845 Wasp Blvd., Building 176, Honolulu, HI, USA*
[13]*Joint Institute for Marine and Atmospheric Research, Pacific Islands Fisheries Science Center, University of Hawaii,Honolulu, HI, USA*
[14]*Instituto de Fomento Pesquero (IFOP), Avda. Blanco Encalada 839, Valparaiso, Chile*
[15]*Fisheries and Oceans Canada, Gulf Fisheries Centre, PO Box 5030, Moncton, NB E1C 9B6, Canada*
[16]*Wageningen UR, Institute for Marine Resources and Ecosystem Studies (IMARES), PO Box 68, 1970 AB Ijmuiden, Netherlands*
[17]*Ifremer, Unité Sciences et Technologies Halieutiques, ZI de la pointe du diable, CS 10070, 29280 Plouzané, France*
[18]*Marine Research Division, AZTI-Tecnalia, Txatxarramendi ugartea z/g, E-48395 Sukarrieta, Bizkaia, Spain*
[19]*IPMA-Instituto Português do Mar e da Atmosfera, I.P. Av. Brasilia, 1449-006 Lisboa, Portugal*
[20]*ICCAT Secretariat, Corazón de María 8, 28002 Madrid, Spain*
[21]*Centro Nacional Patagónico, Blvd. Brown 2915, 9120 Puerto Madryn, Chugut, Argentina*
[22]*CSIRO Wealth from Oceans Flagship, Queensland Biosciences Precinct, 306 Carmody Road, St. Lucia, QLD 4067, Australia*
[23]*International Seafood Sustainability Foundation, 805 15th Street NW, Washington, DC 20005, USA*
[24]*Food and Agriculture Organization of the United Nations, Vialle delle Terme di Caracalla, 00153 Rome, Italy*
[25]*European Commission Joint Research Center, TP 051, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy*
[26]*NOAA NMFS, 17109 Pt. Lena Loop Road, Juneau, AK, USA*
[27]*Integrated Statistics, 172 Shearwater Way, Falmouth, MA, USA*

**Corresponding author: Research Fishery Biologist, NOAA NMFS, 166 Water Street, Woods Hole, MA 02543, USA. tel: +1 508 495 2310; fax: +1 508 495 2393; e-mail: jonathan.deroba@noaa.gov*

†Deceased.

Deroba, J.J., Butterworth, D.S., Methot, R.D. Jr., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J., Valero, J.L., Needle, C.L., O'Malley, J.M., Chang, Y-J., Thompson, G.G., Canales, C., Swain, D.P., Miller, D.C.M., Hintzen, N.T., Bertignac, M., Ibaibarriaga, L., Silva, A., Murta, A., Kell, L.T., de Moor, C.L., Parma, A.M., Dichmont, C.M., Restrepo, V.R., Ye, Y., Jardim, E., Spencer, P.D., Hanselman, D.H., Blaylock, J., Mood, M., Hulson, P.-J. F., Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. – ICES Journal of Marine Science, 72: 19–30.

Received 30 August 2013; accepted 11 December 2013; advance access publication 18 January 2014.

The World Conference on Stock Assessment Methods (July 2013) included a workshop on testing assessment methods through simulations. The exercise was made up of two steps applied to datasets from 14 representative fish stocks from around the world. Step 1 involved applying stock assessments to datasets with varying degrees of effort dedicated to optimizing fit. Step 2 was applied to a subset of the stocks and involved characteristics of given model fits being used to generate pseudo-data with error. These pseudo-data were then provided to assessment modellers and fits to the pseudo-data provided consistency checks within (self-tests) and among (cross-tests) assessment models. Although trends in biomass were often similar across models, the scaling of absolute biomass was not consistent across models. Similar types of models tended to perform similarly (e.g. age based or production models). Self-testing and cross-testing of models are a useful diagnostic approach, and suggested that estimates in the most recent years of time-series were the least robust. Results from the simulation exercise provide a basis for guidance on future large-scale simulation experiments and demonstrate the need for strategic investments in the evaluation and development of stock assessment methods.

Keywords: cross-test, model comparison, pseudo data, self-test, time-series analysis, vpa.

Introduction

Simulation testing has been suggested to evaluate the ability of assessment models to accurately and precisely estimate stock conditions under a range of scenarios (NRC, 1998; Restrepo, 1998; Punt *et al.*, 2002; Kell *et al.*, 2007). More specifically, this methodology has been used to examine issues associated with data availability, model misspecification (i.e. structural uncertainty), and the effect of observation and process errors (ICES, 2004; Linton and Bence, 2008; Wetzel and Punt, 2011; Deroba and Schueller, 2013). Much of this previous simulation work, however, was based on generic fish populations or was designed for applications to specific fish stocks (Kell *et al.*, 1999; Haltuch *et al.*, 2008). Some have suggested that results from generic studies are too broad to be valid for specific cases, while, conversely, others have argued that specific applications are too narrow to be generally relevant (ICES, 2012a).

A simulation exercise that attempts to address these purported short-comings was developed by the International Council for the Exploration of the Sea's (ICES) Methods Working Group (WGMG) in support of the ICES Strategic Initiative for Stock Assessment Methods (SISAM; ICES, 2012a, b). In the context of rapid proliferation of stock assessment methods, ICES and other Regional Fishery Management Organizations recognized the need for reliable stock assessment methods, and SISAM was designed to assure that scientists can apply the best stock assessment methods when developing management advice for fisheries. The wide range of stock assessment modelling approaches being applied were categorized and a simulation-based evaluation of the performance of stock assessment methods under various conditions was developed.

Simulations could be conducted for a range of specific stocks (i.e. based on real data and model fits to the data) covering a breadth of life history, data availability, and fishery type. The simulation exercise would then produce results tuned to the data for each stock, but patterns of results among stocks would also allow for generic advice and conclusions (ICES, 2012a). Results from the SISAM exercise were reported and discussed during a 2-day workshop and 3-day symposium as part of the World Conference on Stock Assessment Methods (WCSAM) in Boston, MA, USA, in July 2013. WCSAM

Table 1 Stocks for which real datasets were used in the simulation exercise

Common name	Scientific name	Assessment model challenges
North Sea cod	*Gadus morhua*	Unallocated removals, variable natural mortality
North Sea plaice (reconstructed discards)	*Pleuronectes platessa*	Shifts in population distribution, subsequent variation in catchability
North Sea plaice	*Pleuronectes platessa*	Discard estimation
North Sea herring	*Clupea harengus*	Internal vs. external stock–recruit estimation, stock structure, variable natural mortality
North Sea haddock	*Melanogrammus aeglefinus*	Time varying selectivity, stock structure, recruitment pulses
Northern hake	*Merluccius merluccius*	Dome selectivity, truncated age structure
Spurdog	*Squalus acanthias*	Sexual dimorphism
Bay of Biscay anchovy	*Engraulis encrasicolus*	Short-lived, high and variable natural mortality
Iberian sardine	*Sardina pilchardus*	Dome selectivity
Southern horse mackerel	*Trachurus trachurus*	Survey year effects, time varying selectivity
North Atlantic albacore tuna	*Thunnus alalunga*	Unknown selectivity and catchability, uncertain growth and natural mortality
US west coast canary rockfish	*Sebastes pinniger*	Dome selectivity, lack of contrast, ageing error, uncertain stock–recruitment
Georges Bank yellowtail flounder	*Limanda ferruginea*	Retrospective pattern
South African anchovy	*Engraulis encrasicolus*	Uninformative age data, uncertain natural mortality

brought together stock assessment scientists from around the world and was uniquely suited for conducting, presenting, and discussing the proposed simulation exercise. The global level of participation broadened the applicability of the results.

This article summarizes a portion of the simulation exercises completed for the SISAM meeting. It also attempts to focus on results that are of broad interest, so that the details for specific stocks are used as examples in support of general conclusions. Because a global simulation exercise like SISAM had never been previously conducted, but may continue in the future, the discussion also highlights the positive features of the process as well as recommendations for improvements and continuing the research.

Methods

The SISAM Steering Committee with support of the ICES WGMG selected 14 datasets from stocks intended to cover a range of life histories, data availabilities, and assessment challenges for consideration in the SISAM exercise (Table 1; ICES, 2012a). Throughout this article, the term "real" data refers to the actual data used for conducting a stock assessment and providing management advice for a given stock, while the term "pseudo-data" refers to computer-generated observations with the characteristics detailed below. The models currently used to assess the stocks and to provide management advice include delay-difference models, age-based models, length-based models, as well as age- and length-based models (ICES, 2012a). The details of the assessment models used in the SISAM exercise were not provided here in the interest of brevity and because these details were not necessary for understanding the results or conclusions.

The research covered in this article is a portion of that conducted for the SISAM exercise and was presented here as two steps (ICES, 2012a). For Step 1, the 14 real datasets were made available to stock assessment scientists throughout the world and the assessment model of their choice was fit using these real data. At this step, the assessment model was not necessarily optimized and may have contained issues of fit that in practice might require additional refinement. Most participants, however, suggested that assessment model fits were optimized (e.g. considered residuals, used measures of statistical fit), as might be done in an actual assessment. In Step 2, characteristics of the fit of a model to real data from Step 1 were used to generate pseudo-data with error. The pseudo-data were then made available to the participants and assessment models were fit to each simulated pseudo-dataset.

Most assessment models assumed that both observation and process errors were present. No clear distinction about the source of the errors in the pseudo-data was made here, however, because it was often not possible to isolate the sources of error from a given assessment model fit. The pseudo-data for each assessment model fit to real data were generated by adding errors to the values for abundance indices, age composition (i.e. from surveys or catches), and total catch. Some assessment models assume that catches are observed without error (e.g. a virtual population analysis) so that catch pseudo-data were generated without errors in these cases. The values from the assessment model fit to real data, however, were based on a set of parameters, process errors, and observation errors that were specific to each model.

Table 2 The number of assessment models applied to real or pseudo-data using different modelling frameworks or different structural assumptions within the same modelling frameworks for four assessment model types

Assessment model type	Number of variants applied
Delay difference	1
Virtual population analysis	4
Statistical catch at age	21
Surplus production	4

Pseudo-data generation with error

For each fit of a stock assessment model to real data, 100 sets of pseudo-data with error were generated using the age-based Population Simulator (PopSim; NOAA, 2013). An age- and length-based variant of PopSim has been used previously to evaluate the performance of stock assessments when natural mortality is misspecified (Deroba and Schueller, 2013). The version used for SISAM, however, was strictly age-based. Owing to the age-based restriction of PopSim, pseudo-data were able to be generated only for assessment models that were also strictly age-based. Several other simulation exercises not using PopSim were carried out for the SISAM workshop, but are not covered in this article.

PopSim required several inputs that were held constant among each of the pseudo-datasets: annual fully recruited fishing mortality rate, annual fishery selectivity at age, population size at age in the first year of the simulation, annual recruitment, annual mean weights-at-age associated with harvested and spawning fish, annual maturity at age, annual natural mortality at age, annual survey catchability, and annual survey selectivity at age. The inputs above allow the simulation of a population that matches the estimated population from the fit of an age-structured assessment model. Without errors, the pseudo- survey observations, age compositions, and catches would also match those of the predicted values from a given assessment fit.

Errors in the pseudo-data for annual indices of abundance and the total fishery catch had a lognormal distribution (Appendix). The degree of variance in the errors was based on the residuals of the fit of a given assessment model to the real survey or catch observations and was constant among years. For assessment models that assumed catch was known without error (e.g. virtual population analysis), the variance of the errors in the catch equalled zero. The errors were assumed to be independent and identically distributed, which was consistent with most, but not all the assessment models considered.

Table 3 Number of assessment model fits to the real datasets for each stock

Species	Models fit to real data
George's Bank yellowtail flounder	13
North Sea cod	11
North Sea herring	10
Southern horse mackerel	6
Iberian sardine	5
North Sea haddock	5
Spurdog	3
South African anchovy	2
US west coast canary rockfish	2
Bay of Biscay anchovy	1
Northern hake	1
North Sea plaice (reconstructed discards)	1
Total	60

Pseudo-data of survey and catch age compositions with were generated by drawing annual samples from a multinomial distribution (Appendix). Annual sample sizes for the multinomial distribution were either the same as those assumed in fitting the assessment model to real data (i.e. when the assessment model assumed a multinomial distribution) or were based on input from the scientist who conducted the assessment model fit (i.e. when the assessment model did not assume a multinomial distribution). For assessment models that assumed age composition was known without error (e.g. virtual population analysis), the annual sample sizes were set at a relatively high value. Thus, the extent of error in the pseudo-data was generally consistent with the assessment model fit to real data.

Output metrics

For assessment model fits to real and pseudo-data, time-series estimates of stock biomass (spawning stock in most cases, total biomass in others) and fishing mortality (fully selected in most cases, an average among ages in others) were recorded. For assessment model fits to real data from each stock, the time-series estimates were plotted and qualitatively examined for variation in scale and trend. Because the measures of biomass and fishing mortality were not standardized among all assessment model fits, each time series was also rescaled to have a mean of 1.0 by dividing each time series by its average. A plot of the rescaled time series partially accounted for the issue of inconsistent measures and was easier to compare for variation in temporal trends.

The fits of assessment models to pseudo-data with error allowed for two general types of comparisons, which were termed self-tests and cross-tests. In self-tests, an assessment model was fit to the 100 pseudo-datasets generated in a manner consistent with the fit of the same assessment model (i.e. same modelling platform with the same structural assumptions and model settings) to real data for a given stock. Summary statistics (i.e. median, 10th percentile, 90th percentile) of the fits to the 100 pseudo-datasets were then plotted with results of the fit to the real data. The time-series estimates from the fit to real data provided a basis of comparison for the fits from the 100 pseudo-datasets. The term self-test was used because the assessment model fit to the real data and to the pseudo-data were identical (i.e. same assessment platform, structural assumptions, and settings). In cross-tests, an assessment model was fit to 100 pseudo-datasets generated in a manner consistent with the fit of a different assessment model (possibly same assessment platform with different structural assumptions or model settings) to the real data for a given stock. Summary statistics (i.e. median, 10th percentile, 90th percentile) of the fits to the 100 pseudo-datasets from a given assessment were then plotted together with the fit of the alternative assessment model to the real data. The time-series estimates of the alternative assessment model fit to real data provided a basis of comparison for the fits from the 100 pseudo-datasets. The term cross-test was used because the assessment model fit to the real data and to the pseudo-data were different (i.e. different assessment platforms, different structural assumptions, or different model settings). Plots were qualitatively examined for

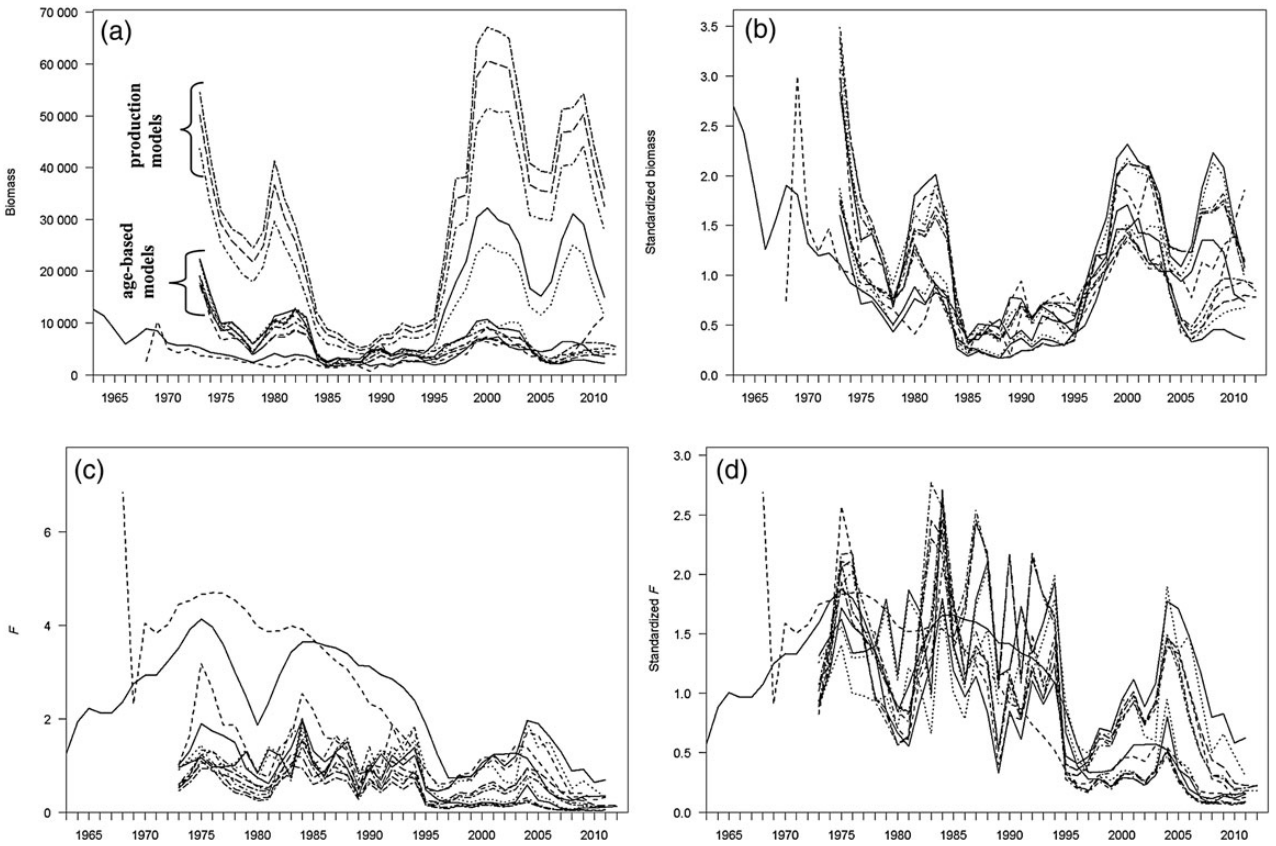

Figure 1. Time-series estimates of biomass and fishing mortality from the assessment model fits to real data for Georges Bank yellowtail flounder on absolute (a and c) and standardized scales (b and d). The different line types denote different assessment model fits and are only intended to help with tracking individual time series.

consistency between the time series based on the fit to real data and the summary statistics from the fits to 100 pseudo-datasets. Conditional on a given assessment model fit and the errors added (e.g. underlying statistical distributions for errors), divergence between the time series from the fit to real data and the summary statistics suggested a lack of robustness. Lack of robustness in self-tests may be indicative of bias, whereas lack of robustness in cross-tests might be expected due to differences between models (e.g. structural assumptions, statistical assumptions, etc.).

Results

Thirty different assessment models were applied at some step of the SISAM exercise, either using different modelling frameworks or different structural assumptions within the same modelling framework (Table 2). Most assessment models were statistical catch at age, while the remainder were virtual population analysis, surplus production, or delay difference models (Table 2).

Assessment models were fit to real data in 60 unique combinations covering 12 of the 14 datasets selected by the WGMG (Table 3). The extent of variation among assessment model runs depended upon the models applied. Greater variation in scale was generally evident when surplus production models that produced time series of fishable biomass were applied to the real data along with other assessment models that could broadly be classified as age-based assessment models that provided spawning-stock biomass. For example, the extent of variation in the scale of biomass estimates was relatively large for Georges Bank yellowtail flounder (Figure 1a and c). Conversely, the North Sea herring stock, to which only statistical catch-at-age-type models were applied, exhibited an extent of variation in the scale of the time-series estimates that was relatively small (Figure 2a and c).

Time-series trends among assessment model fits to real data for each stock were generally similar. The extent of variation in temporal trends among assessment model fits, however, depended upon characteristics of the real data for each stock, with generally less variation for those stocks with broadly consistent real data and greater variation and different temporal trends among fits for those stocks with known inconsistencies in the real data. For example, real data for North Sea herring are relatively consistent, with periods of high and low abundance (i.e. high contrast), and the temporal trends in the rescaled time series were similar (Figure 2b and d). Results for North Sea cod and spurdog were also relatively consistent (Figures 3b, d, and 4b and d). Conversely, the temporal trends of the rescaled time series among assessment model fits to real data for Georges Bank yellowtail flounder manifest several inconsistencies (Figure 1b and d). Recent decreases in fishing mortality on Georges Bank yellowtail flounder, however, have been accompanied by increases in indices of abundance, but no expansion of the age structure (i.e. no increase in the proportion of relatively older fish in the population; Legault *et al.*, 2012). This pattern in the data was resolved among the assessment model runs in a variety of ways, mostly through allowing some parameter to vary over time (including natural mortality, survey catchability, intrinsic growth rate, and selectivity). These different solutions to the patterns in the data produced different time series trends (Figure 1b and d). Results for southern horse mackerel were also relatively inconsistent, which may be related to year effects and poor internal consistency apparent in bottom trawl survey data (ICES, 2011; Figure 5b and d).

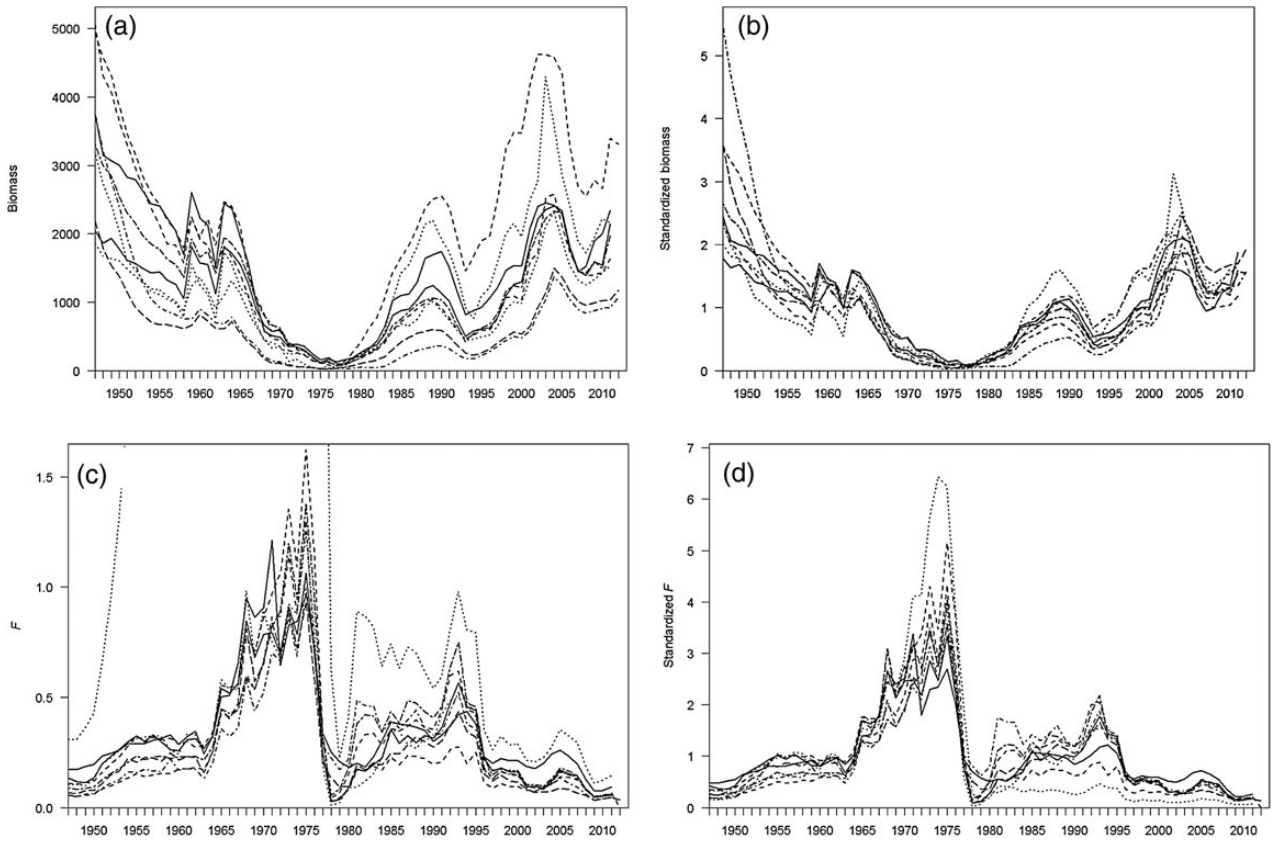

Figure 2. As in Figure 1 except for North Sea herring.

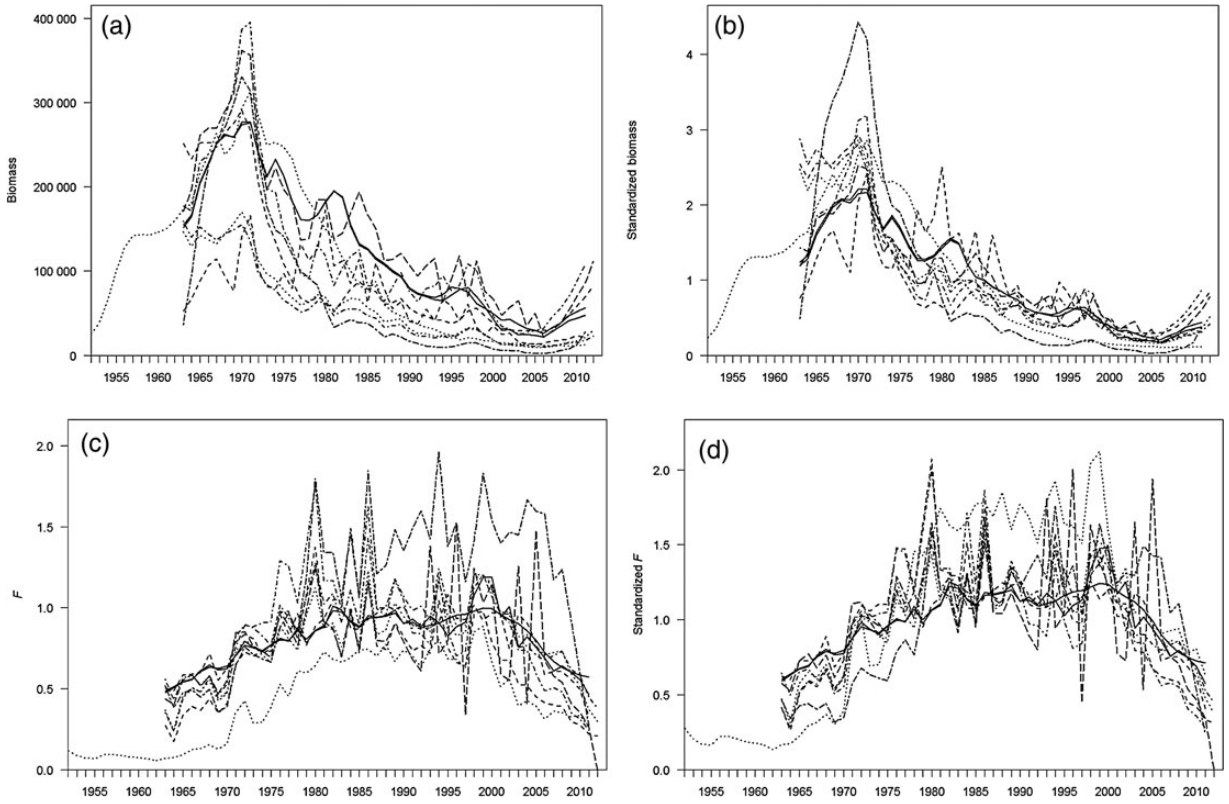

Figure 3. As in Figure 1 except for North Sea cod.

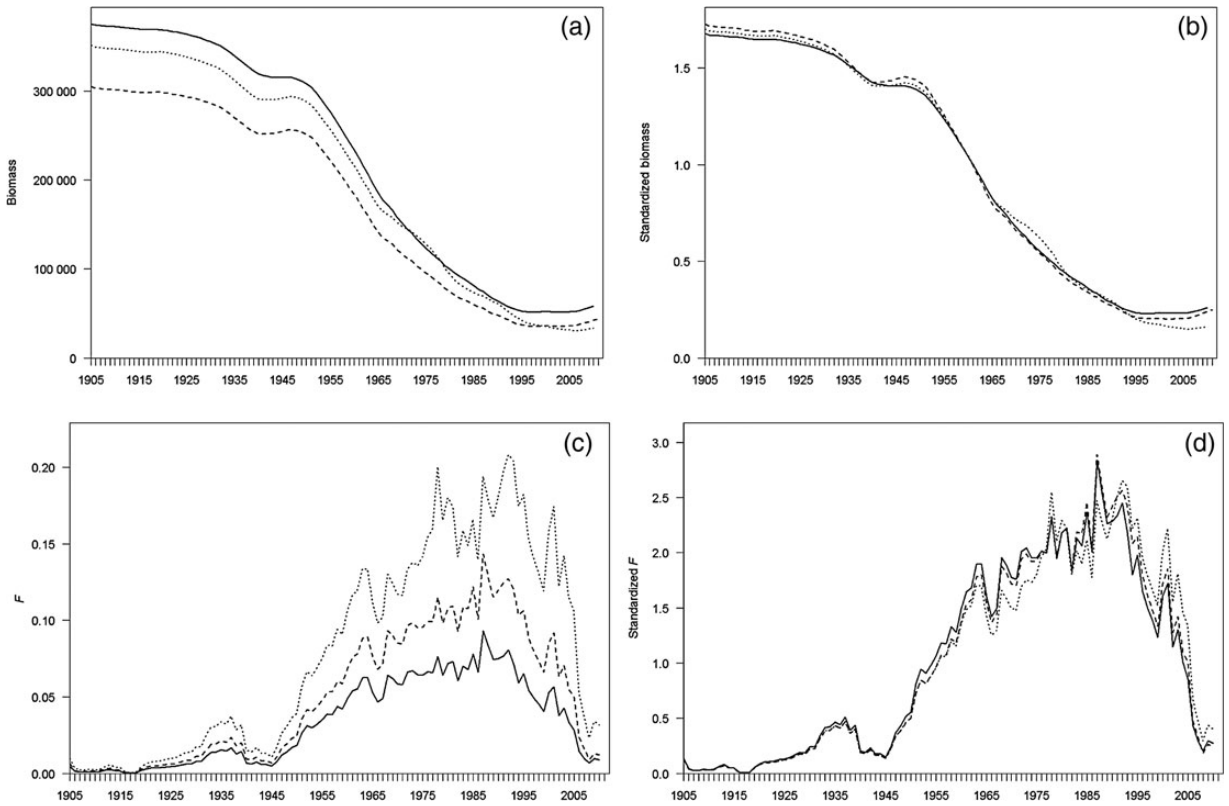

Figure 4. As in Figure 1 except for spurdog.

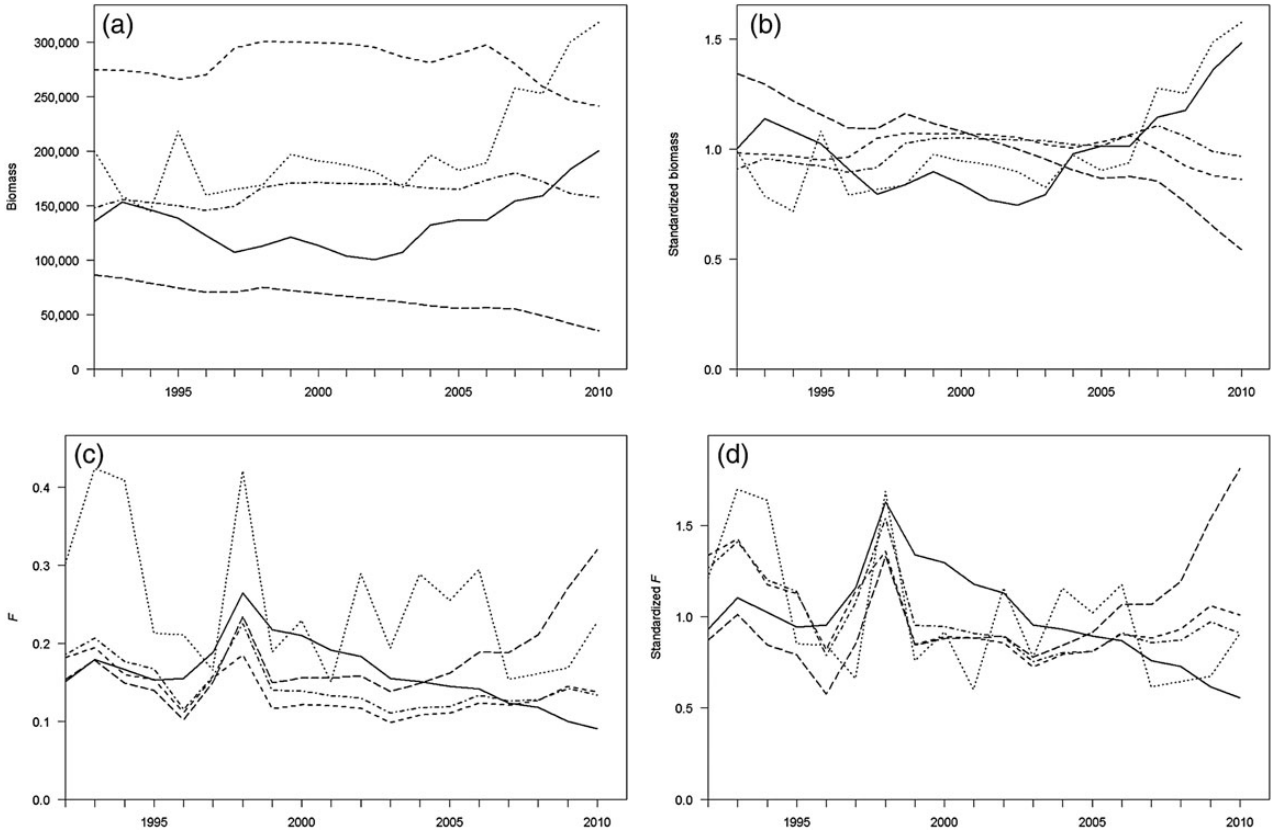

Figure 5. As in Figure 1 except for southern horse mackerel.

Table 4 Number of self- and cross-tests completed by stock

Species	Self-tests	Cross-tests
George's Bank yellowtail flounder	3	3
Iberian sardine	2	0
North Sea cod	1	2
Southern horse mackerel	1	0
Total	7	5

Assessment model self-tests were conducted on seven unique combinations covering four stocks (Table 4). Five of the seven self-tests showed some divergence and always in the more recent years. The extent of divergence in the self-tests depended on the stock and assessment model. For example, self-tests for Georges Bank yellowtail flounder had relatively worse divergence than self-tests for Iberian sardine (Figures 6 and 7). Furthermore, for Georges Bank yellowtail flounder, the divergence in the self-test of a virtual population analysis with a random walk in natural mortality was worse than a statistical catch-at-age model with a random walk in catchability (Figure 6). Most self-tests used assessment models that allowed for a parameter to follow a random walk and it is not clear whether differences in performance are driven by basic model type (i.e. virtual population analysis or statistical catch at age) or the different hypotheses about time-varying parameters.

Assessment model cross-tests were conducted on five unique combinations covering two stocks (Table 4). All cross-tests showed some divergence, and similar to self-tests, the divergence occurred in more recent years in all but one case. For example, divergence was the worst or nearly worst in the time series for Georges Bank yellowtail flounder and North Sea cod cross-tests in the most recent years (Figures 8 and 9). Similar to self-tests, all cross-tests used assessment models that allowed for a parameter to follow a random walk. Divergence in biomass estimates in cross-tests, however, was less when the same parameter was allowed to follow a random walk in the model fit to the real data and the model fit to the pseudo-data than when the random walk parameters differed between models (Figure 8).

Discussion

Although not explicitly examined, the variation in scale and trend from fits to real data by different assessment models seems comparable to or greater than measures of within model variation (from, for example, Markov Chain Monte Carlo simulation, bootstrapping) for some stocks, as others have found (Ralston *et al.*, 2011). This variation, which likely represents structural uncertainty to a large extent, should be presented to managers regularly in addition to the usual measures of within model variation (e.g. parameter uncertainty) so that a more accurate representation of uncertainty can be considered, as has been suggested by others (Williams, 1997; SEDAR, 2010). Although presenting the extent of among-model variation to managers may convey uncertainty more accurately, such an approach is likely to complicate management decisions (e.g. setting annual quotas) that often rely on having a single, "best" assessment model. Methods for making multi-model inference (e.g. model averaging; Brodziak *et al.*, 2015) is an area of active research, however, which should continue in the future (Brodziak and Legault, 2005; Anderson, 2008). In a real assessment

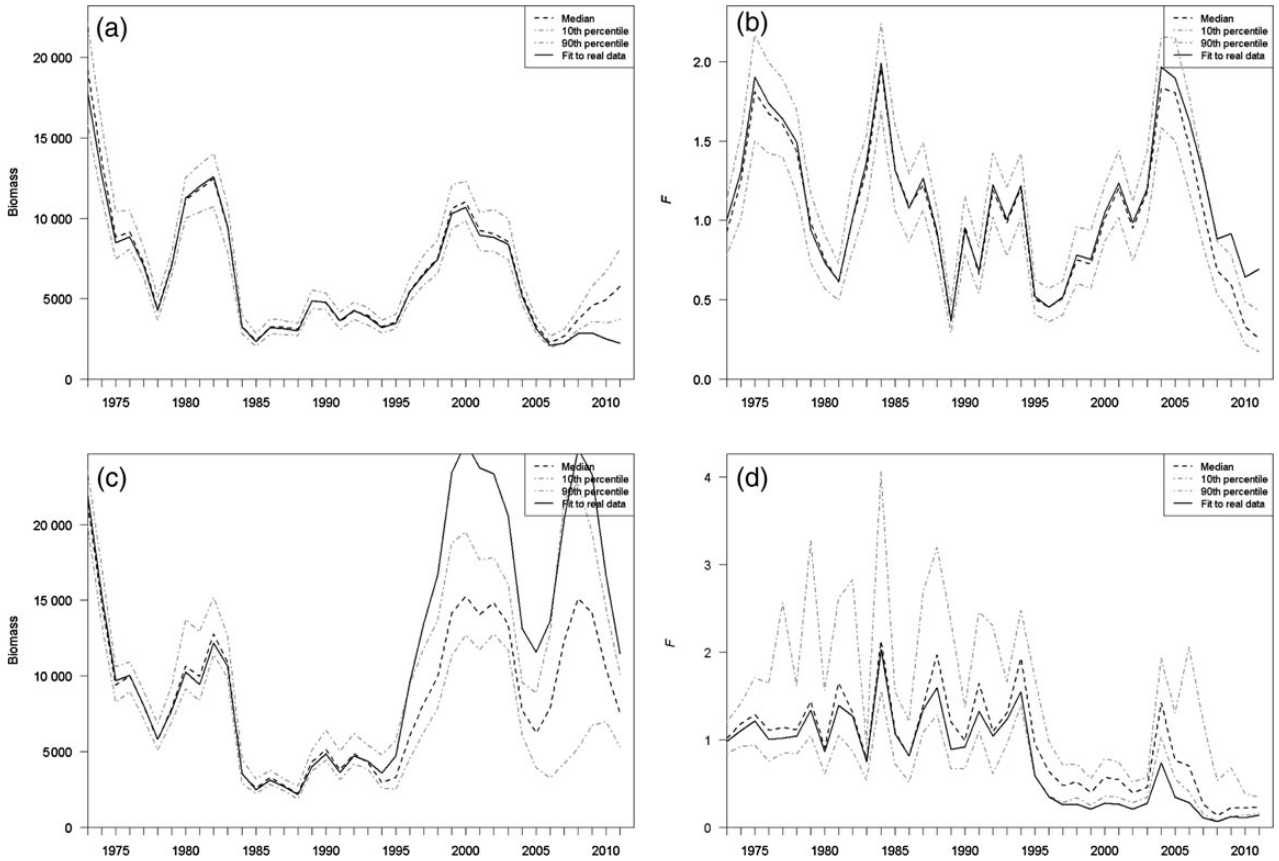

Figure 6. Results of self-tests for Georges Bank yellowtail flounder for a statistical catch-at-age model with a random walk in catchability (a and b) and a virtual population analysis with a random walk in natural mortality (c and d). The same assessment model with identical settings was used in the fit to real data as in the fits to the pseudo-datasets summarized as percentiles in the graphs.

situation, considerable attention is paid to the exact setup of each model considered so that some of the among-model variation evident here might be reduced. The finding of among-model variation is an indication of the need for better business practices for setup of all assessment models so that they can be applied in as consistent a manner as possible. Likewise, management procedures (management strategy evaluation) are designed to test and recommend management actions and assessment methods that take into account parameter and structural uncertainty, and such work should also continue (Butterworth and Punt, 1999; Butterworth, 2007). A formal comparison of the relative magnitude of among- and within-model variations was beyond the scope of this article, but should also be a topic of future research. Ignoring among-model variation, as is commonly the case, is likely to result in an under-representation of stock assessment uncertainty (Magnusson *et al.*, 2012).

Divergence in self-tests was common and this type of consistency check should likely be conducted whenever possible (e.g. as assessment frequency or workload allow) and become a standard method when considering an assessment model for management advice (Piner *et al.*, 2011). Institutions involved in scientific advice should add such tests to their best practices or guidelines for long-term methods development. When divergence in self-tests occurs, the inconsistency should be investigated for an explanation. A first step would be to audit code, and model settings to ensure the behaviour of the data-generating model (commonly called the "operating model") and -assessment model (commonly called the "estimation model") are consistent with expectations. In cases where the estimation model is not entirely consistent with the data-generating model (e.g. estimation model is a simplification of the data-generating model), as was common in this manuscript, another follow-up exercise would be to apply the assessment model to pseudo-data generated without errors. Continued divergence would suggest that the cause is structural uncertainty, while consistent results would suggest that the errors themselves are the cause. These types of follow-up exercises have not yet been conducted for results of the SISAM workshop and symposium due to the scale of the undertaking, but are planned for the future.

Divergence in cross-tests was also common. The assessment models with divergent results in cross-tests during SISAM also had divergent results in self-tests. Divergence in self-tests had generally been unexpected, and understanding the reasons for such lack of robustness in self-tests should take priority over understanding reasons for divergence in cross-tests. Thus, explanations for divergence in cross-tests of the SISAM exercise have not yet been explored. A possible way to conduct such explorations would be to fit the assessment model to the pseudo-data, but to impose a penalty on the likelihood for deviations from the time series based on the fit to real data. A comparison of the likelihood component values for the assessment model with and without the penalty should reveal the data source or parameters largely responsible for inducing the divergence. Furthermore, initially conducting cross-tests

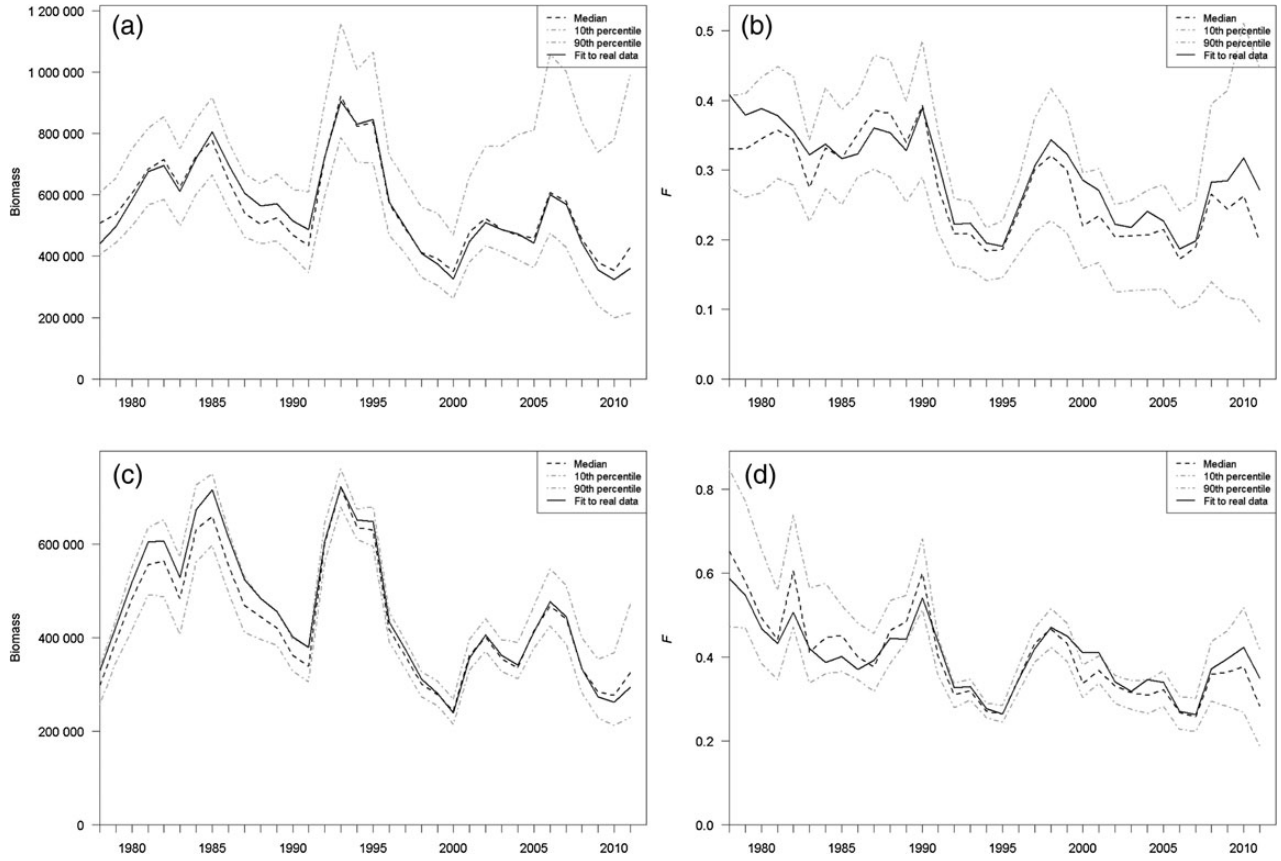

Figure 7. As in Figure 6 except for Iberian sardine and for an age-based assessment model (a and b) and an integrated analysis model that is also age-based (c and d).

using the same modelling framework but with two different sets of structural assumptions or model settings might be informative and more tractable because the chances for divergence induced by unforeseen structural uncertainties would be diminished and allow the analyst greater control.

Divergence in self- and cross-tests was especially common for models that featured temporal random walks for some parameters (e.g. catchability, natural mortality). This result is counter to some previous research that found allowing random walks was rarely detrimental and often reduced bias (e.g. Wilberg and Bence, 2006). Most of the assessment model applications used for self- and cross-tests, however, had a random walk parameter and few applications were available for comparison that did not feature these time-varying dynamics. This topic, however, likely warrants further consideration.

This manuscript has examined time-series estimates of stock biomass and fishing mortality, but a range of other output metrics could be considered. For example, during the SISAM exercise, requests were also made for reference points related to stock biomass, fishing mortality, and yield. Using reference points would permit the relative stock status among assessment models to be compared directly and eliminate some issues of scale. Difficulties were encountered, however, in standardizing the reference point and methods to be used for estimating relative stock status, and furthermore the reference point being requested varied by stock. For example, some models (e.g. production models) can provide direct estimates of common reference points such as maximum sustainable yield, but other models cannot or require some calculations external to the assessment fit (e.g. some age-based models). Some difficulties were also likely caused by variations in the way reference points are utilized by international management bodies. Consequently, the utility and general interest of dedicating time to calculating some reference points varied by participant. Several solutions could likely be applied in future research. For example, the reference point and method of reference point calculation could be standardized among all stocks and assessment models, which might permit code sharing and reduce time commitments. Alternatively, a relatively simple metric could be used, such as biomass in the final year of the assessment divided by biomass in the first year or the time-series average. A measure of an assessment model's predictive power might also be a useful metric, especially since parameter estimates in more recent years are typically the least robust, as was common in this study. A measure of forecasting ability could serve as another measure of robustness and provide objective weights for multi-model inference, such as model averaging.

A common objective for simulation testing like that done for SISAM is to evaluate an assessment model's capacity to provide robust catch advice. Thus, bias in the absolute scale of biomass or fishing mortality rate estimates may not be a problem if the resulting advice for a sustainable catch level is unbiased. For example, the bias in the absolute scale of biomass and fishing mortality estimates is often in opposite directions, such that the product of biomass and a desired fishing mortality rate could result in unbiased catch

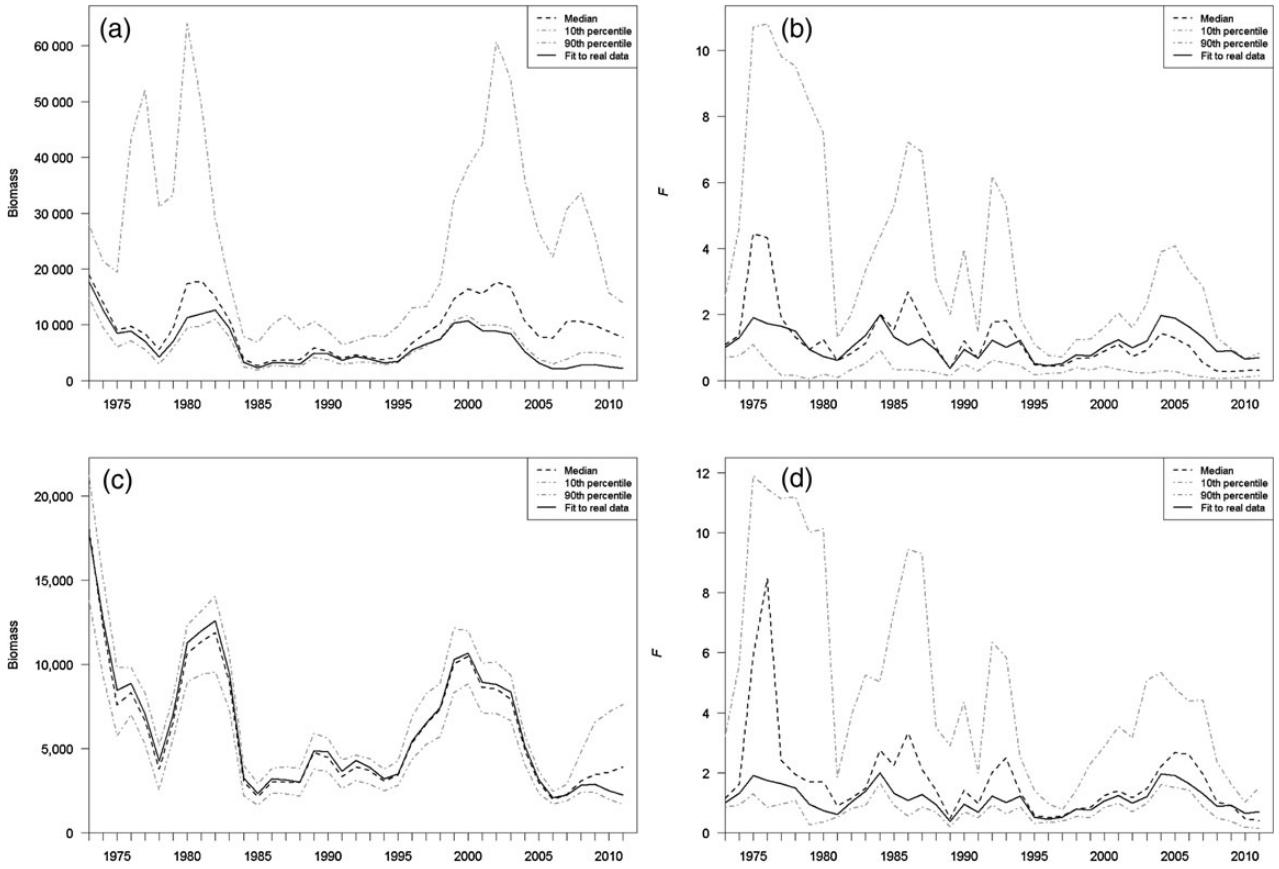

Figure 8. Results of cross-tests for Georges Bank yellowtail flounder. A statistical catch-at-age model with a random walk in catchability was used for the fit to real data in all panels, but fits to pseudo-datasets summarized as percentiles were based on applying a virtual population analysis with a random walk in natural mortality (a and b) or catchability (c and d).

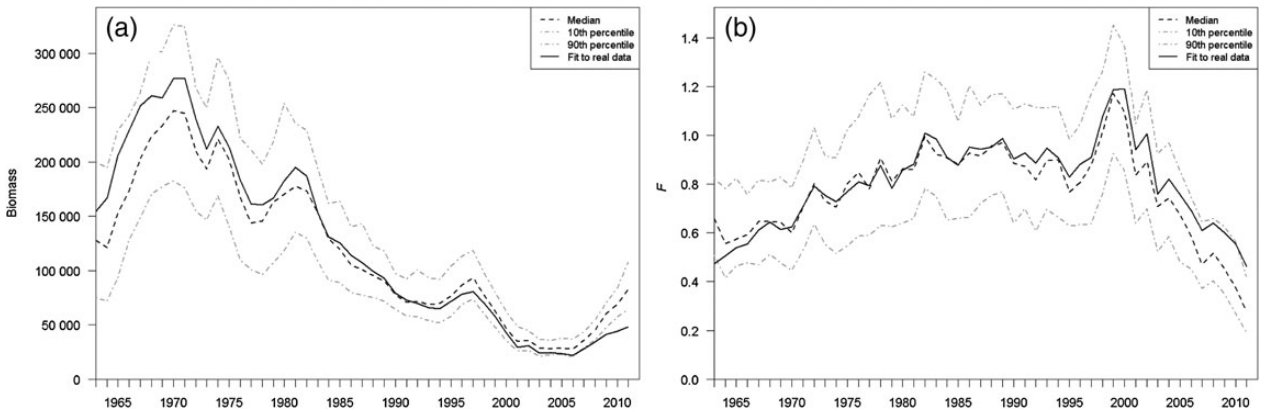

Figure 9. As in Figure 8 except for North Sea cod. Different age-based assessments with different structural assumptions were used for fits to real data and pseudo-datasets.

advice. Therefore, alternative output metrics related to short-term or sustainable catch advice (e.g. constant catch level for the next 5–10 years that would maintain current biomass) should likely be considered in the future. Furthermore, the presence of retrospective patterns (Mohn, 1999) and uncertainty in parameter estimates (e.g. relative estimation error) or stock forecasts could also be evaluated for their effects on the ability of an assessment model to provide robust catch advice.

The assessment model fits and simulations completed for SISAM were designed to provide general guidance on model performance and to introduce assessment model validation techniques that are likely of broad interest. Future research, however, could expand the general SISAM simulation methodology to examine a range of specific topics, including evaluations of sampling programme investments, causes and consequences of retrospective patterns, the extent of complexity of the data-generating model (e.g. sexual

dimorphism, ecosystem models, stock structure, spatial heterogeneity, fleet dynamics), data availability, and degree of data aggregation (e.g. error in age/length composition, tagging data, predatory consumption). Exploring specific topics, however, will likely require continued international collaborations to achieve enough treatments among stocks so that broad conclusions can be drawn, as initiated by SISAM.

SISAM utilized PopSim as a centralized pseudo-data-generating model (i.e. operating model). This approach had both advantages and disadvantages. By using the same pseudo-data-generating platform for all simulations, the type and format of pseudo-data was standardized among all users. Furthermore, the use of a single pseudo-data-generating model did not afford any specific assessment model a clear advantage over any other. Using a single pseudo-data-generating model, however, was inflexible, and PopSim could not sufficiently replicate all features of some assessment models and underlying statistical distributions were not always consistent between a given assessment model and PopSim. Furthermore, a true self-test should have complete consistency, including the underlying statistical assumptions, between the pseudo-data-generating model and the assessment model. By failing to achieve this consistency, true divergence and potential bias could not be distinguished from divergence caused by differences between the estimation model and data-generating model (e.g. underlying statistical distributions) for some self- and cross-tests. Nonetheless, divergence caused by differences between the estimation model and data-generating model still suggests a lack of robustness. Continued simulation research like that of SISAM should carefully weigh the competing objectives of centralized control and equitable treatment against consistency between the pseudo-data-generating and assessment models.

Data from the 14 stocks used in SISAM were made available to all participants, and questions raised by the participants when applying an assessment model to data (real or pseudo-data) from a given stock were resolved by an individual expert, often with help from a single assessment scientist who acted as liaison between participants and stock experts. This approach created an unbalanced workload for the stock experts and liaison, and this often led to delays in resolving questions. Furthermore, in practice a model is often applied to a real dataset for the first time in the context of an extensive review process that includes time for dialogue and repeated model fits to ensure optimized tuning for the given situation. Time constraints prevented a thorough review of model fits during the SISAM exercise so that some results might be driven by suboptimal assessment model settings. While controlling for this issue might be preferred, the possibility of results being driven by suboptimal settings does illustrate the importance of practitioner expert in the assessment process.

An alternative organizational scheme that might alleviate some of the problems above would be to stratify participants by stock, objective to be addressed, or some other feature of interest (e.g. assessment model type). Stock experts could then respond to questions from a subset of participants only, ensure proper interpretation of data (e.g. units), and could be more directly involved with each application of an assessment model to a given dataset. This stratification scheme would have the benefits of more efficient communication and would be less error prone because participants would have better knowledge of a single stock in contrast to limited knowledge of multiple stocks. Some drawbacks, however, would be that participants may not get to research all stocks or questions that interest them, and distilling results into broad conclusions would require coordination among the stratified groups.

In conclusion, the simulation exercise highlighted the following issues:

(i) Different models were consistent in regard to estimating trends, but did not consistently estimate the scale of absolute biomass.

(ii) Similar types of models (age based, production, etc.) generally behaved in similar manners. In other words, the choice of model type had the biggest effect on consistency across models.

(iii) Self-testing is useful and should be encouraged.

(iv) Self- and cross-testing frequently highlighted divergence in the most recent years of time series.

(v) Among model variability can be considered as a type of uncertainty and has implications when considering whether to apply a purpose built "best fit" model or to use an ensemble approach (e.g. model averaging).

These findings demonstrate the value of simulation-based evaluations of model performance. The difficulties experienced in this broad, inclusive process also offer guidance on the conduct of future large-scale simulation exercises. Finally, results suggest that further strategic investments are needed for the advancement of stock assessment methods for supporting management of sustainable fisheries.

Acknowledgements

We are grateful to the participants and support staff of the WCSAM workshop and symposium. We also thank many agencies and stock assessment scientists who were willing to dedicate their data and time to these efforts. Likewise, this research would not have been possible without the work of those responsible for fishery data collection and storage. We also offer this research and associated future endeavours in memoriam of Alberto Murta, who passed away during preparation of this article. This research addresses the good practices in stock assessment modelling programme of the Center for the Advancement of Population Assessment Methodology (CAPAM).

References

Anderson, D. R. 2008. Model Based Inference in the Life Sciences. Springer, New York, New York.

Brodziak, J., and Legault, C. M. 2005. Model averaging to estimate rebuilding targets for overfished stocks. Canadian Journal of Fisheries and Aquatic Sciences, 62: 544–562.

Brodziak, J., O'Malley, J. M., and Chang, Y. J. 2014. Bayesian multimodel inference for stock assessment. ICES Journal of Marine Science, this issue.

Butterworth, D. S. 2007. Why a management procedure approach? Some positives and negatives. ICES Journal of Marine Science, 64: 613–617.

Butterworth, D. S., and Punt, A. E. 1999. Experiences in the evaluation and implementation of management procedures. ICES Journal of Marine Science, 56: 985–998.

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

Haltuch, M. A., Punt, A. E., and Dorn, M. W. 2008. Evaluating alternative estimators of fishery management reference points. Fisheries Research, 94: 210–303.

ICES. 2004. Report of the Working Group on Methods of Fish Stock Assessments. 11–18 February 2004, Lisbon, Portugal. ICES CM 2004/D:03. 232pp.

ICES. 2011. Report of the Working Group on Methods of Fish Stock Assessment (WGMG), 10–19 October 2011. Vigo, Spain. ICES CM 2011/SSGSUE:08. 250pp.

ICES. 2012a. Working Group on Methods of Fish Stock Assessments (WGMG), 8–12 October 2012. Lisbon, Portugal. ICES CM 2012/SSGSUE:09. 249pp.

ICES. 2012b. Report on the Classification of Stock Assessment Methods developed by SISAM. ICES CM 2012/ACOM/SCICOM:01. 15 pp.

Kell, L. T., Mosqueira, I., Grosjean, P., Fromentin, J. M., Garcia, D., Hillary, R., Jardim, E., *et al.* 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES Journal of Marine Science, 64: 640–646.

Kell, L. T., O'Brien, C. M., Smith, M. T., Stokes, T. K., and Rackham, B. D. 1999. An evaluation of management procedures for implementing a precautionary approach in the ICES context for North Sea plaice. ICES Journal of Marine Science, 56: 834–845.

Legault, C. M., Alade, L., Stone, H. H., and Gross, W. E. 2012. Stock assessment of Georges Bank yellowtail flounder for 2012. TRAC Reference Document 2012/02; 133p. http://www2.mar.dfo-mpo.gc.ca/science/TRAC/rd.html.

Linton, B. C., and Bence, J. R. 2008. Evaluating methods for estimating process and observation error variances in statistical catch-at-age analysis. Fisheries Research, 94: 26–35.

Magnusson, A., Punt, A. E., and Hilborn, R. 2012. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish and Fisheries, 14: 325–342.

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.

National Oceanic and Atmospheric Administration (NOAA). 2013. NOAA Fisheries Toolbox, Version 3.1. Age Based Population Simulator, Version 1.0. http://nft.nefsc.noaa.gov/ (accessed 02.26.13).

National Research Council (NRC). 1998. Improving Fish Stock Assessments. The National Academies Press, Washington, DC.

Piner, K., Lee, H-H., Maunder, M., and Methot, R. 2011. A simulation-based method to determine model misspecification: examples using natural mortality and population dynamics models. Marine and Coastal Fisheries, 3: 336–343.

Punt, A. E., Smith, A. D. M., and Cui, G. 2002. Evaluation of management tools for Australia's South East Fishery. 2. How well can management quantities be estimated? Marine and Freshwater Research, 53: 631–644.

Ralston, S., Punt, A. E., Hamel, O. S., DeVore, J. D., and Conser, R. J. 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fishery Bulletin, 109: 217–231.

Restrepo, V. R. (Ed.) 1998. Analyses of simulated data sets in support of the NRC study on stock assessment methods. NOAA Technical Memorandum, NMFS-F/xSPO 30.

Southeast Data Assessment and Review (SEDAR). 2010. Characterizing and presenting assessment uncertainty. SEDAR Procedural Workshop IV. Ed. by J. Carmichael, pp. 78. http://www.sefsc.noaa.gov/sedar/.

Wetzel, C. R., and Punt, A. E. 2011. Performance of a fisheries catch-at-age model (Stock Synthesis) in data-limited situations. Marine and Freshwater Research, 62: 927–936.

Wilberg, M. J., and Bence, J. R. 2006. Performance of time-varying catchability estimators in statistical catch-at-age analysis. Canadian Journal of Fisheries and Aquatic Sciences, 63: 2275–2285.

Williams, B. K. 1997. Approaches to the management of waterfowl under uncertainty. Wildlife Society Bulletin, 25: 714–720.

Appendix

Details of pseudo-data generation

Errors in the annual indices of abundance, I, had a lognormal distribution:

$$I_y = \sum_a q_y S_{y,a} N_{y,a} e^{\varepsilon_y}; \quad \varepsilon_y \sim N(0, \sigma_I^2);$$

where a was age, y was year, q was catchability, S was selectivity, and N was stock abundance on 1 January of each year. The degree of variance, σ_I^2, in the errors was approximated using a coefficient of variation, CV, which was based on the residuals of the fit of a given assessment model to the real survey observations and was constant among years:

$$\sigma^2 = \ln(CV^2 + 1).$$

Errors in annual total fishery catch, C, had a lognormal distribution:

$$C_y = \sum_a \frac{F_{y,a}}{Z_{y,a}} N_{y,a} W_{y,a} (1 - e^{-Z_{y,a}}) e^{\delta_y}; \quad \delta_y \sim N(0, \ \sigma_C^2);$$

where F was fishing mortality and the values for each year and age were the product of fully selected fishing mortality and fishery selectivity at age, Z was total mortality and the values for each year and age were the sum of $F_{y,a}$ and annual natural mortality at age, and W was the mean weight of a harvested fish. The degree of variance, σ_C^2, in the errors was approximated using a coefficient of variation, CV, which was based on the residuals of the fit of a given assessment model to the real catch observations. The CV was converted to a variance as for indices of abundance. For assessment models, such as virtual population analysis, that assumed catch was known without error, the variance of the errors equalled zero.

Pseudo-data of survey and catch age compositions with error were generated by drawing annual samples from a multinomial distribution. In cases where a given assessment model also assumed that age composition data had a multinomial distribution, the annual sample sizes were the same as those assumed in fitting the assessment model to the real data. In cases where a given assessment model did not assume multinomial distributions for the age compositions, the annual sample sizes were based on input from the scientist who conducted the assessment model fit to the real data. For assessment models that assumed age composition was known without error (e.g. virtual population analysis), the annual sample sizes were set at a relatively high value so that errors in the age composition were negligible. So, in some cases, the underlying statistical distributions that were used to generate pseudo-data were not consistent with the distributions assumed in a given assessment model, but the extent of error in the data was generally consistent with the assessment model fit to real data. The annual proportions at age for the multinomial distribution equalled the expected proportions at age for the given survey or fishery. This process was analogous to disaggregating I_y and C_y using the proportions from the pseudo-age compositions.

Handling editor: Jörn Schmidt