ICES WGMG REPORT 2009

ICES RESOURCE MANAGEMENT COMMITTEE

ICES CM 2009/RMC:12

REF. ACOM, SCICOM

Report of the Working Group on Methods of Fish Stock Assessment (WGMG)

20–29 October 2009

Nantes, France

ICES International Council for the Exploration of the Sea

CIEM Conseil International pour l'Exploration de la Mer

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

H. C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

Recommended format for purposes of citation:

ICES. 2009 . Report of the Working Group on Methods of Fish Stock Assessment (WGMG), 20–29 October 2009, Nantes, France. ICES CM 2009/RMC:12. 85 pp.
https://doi.org/10.17895/ices.pub.9045

Contents

Executive summary **1**

1 Introduction **3**

1.1 Terms of Reference (ToRs) 3

1.2 Report structure 3

2 Working papers **4**

2.1 WP1 – Benoit Mesnil: A history of shrinkage in ICES 4

2.1.1 Abstract 4
2.1.2 Summary of discussion 5

2.2 WP 2 - Anders Nielsen: Status of the state-space assessment model 6

2.2.1 Abstract 6
2.2.2 Summary of discussion 7

2.3 WP 3 - José de Oliveira: Exploratory assessment model for Northeast Atlantic spurdog 7

2.3.1 Abstract 7
2.3.2 Summary of discussion 8

2.4 WP 4 – Coby Needle: XSA convergence 8

2.4.1 Abstract 8
2.4.2 Summary of discussion 9

2.5 WP 5 – Lionel Pawlowski: Overview of the Roundnose Grenadier stock assessment in ICES Vb, VI, VII and XIIb 9

2.5.1 Abstract 9
2.5.2 Summary of discussion 10

2.6 WP 6 – Noel Cadigan: Updates on Noel's version of SURBA 10

2.6.1 Abstract 10
2.6.2 Summary of discussion 10

3 Length-based assessment methods (ToR a1) **11**

3.1 Developments on length-based assessment methods (ToR a1) 11

3.2 Sensitivity of spurdog assumptions about pre-1980 catch-at-age structure 14

4 XSA shrinkage (ToR a4) **19**

4.1 Year-shrinkage 19

4.2 Age-shrinkage 20

4.3 Recommendations 21

5 XSA iteration convergence (ToR a5) **22**

5.1 Introducion 22

5.2 Previous advice on XSA convergence 22

5.3 ICES Working Group approaches to XSA convergence 23

5.4 Convergence tests with ICES Working Group data 25

5.5 XSA convergence for North Sea haddock: comparing XSA and SAM runs 29
5.5.1 Introduction 29
5.5.2 Results 29
5.6 Convergence tests with simulated data 31
5.7 Conclusions 36

6 Review of environmental information in assessments and advice (ToR a6) 38

7 Influence of uncertainty in age–length keys (ToR a7) 41
7.1 Origin and issues with age data 41
7.2 Quantification of uncertainties in the assessment 42
7.3 Effects of the size of the ALK on the assessment 45
7.4 Recommendations 47

8 State-space assessment models 48
8.1 Setting zero variances in a state-space stock assessment model 48

9 Incorporation of survey variance in assessments 52
9.1 Further Extensions to the SURBA model (SURBA+) 52
9.1.1 SURBA+ software 53
9.1.2 3Ps cod example – the survey data 53
9.1.3 Model 1: Random walk for fully recruited fishing mortality 56
9.1.4 Model 2: Trend in Selectivity 59
9.1.5 Model 3: Random Recruitment 62
9.1.6 Recommendations 63
9.2 Developments in SURBA-R 63
9.3 SURBA 2.1 64

10 Conclusions 64
10.1 New chair proposal 64
10.2 Future directions for WGMG 64

11 References 66

Annex 1: List of participants 72

Annex 2: WGMG Terms of Reference for the next meeting 73

Annex 3: Recommendations 79

Annex 4: Stock simulation for testing XSA iteration convergence 80

Executive summary

XSA shrinkage

Shrinkage (either by year or by age) is a relatively *ad hoc* device that was implemented in the XSA model to try to reduce unwanted assessment fluctuations driven by noise rather than signal. We summarize the history of shrinkage in XSA and consider how shrinkage is being used in current ICES assessment working groups. We conclude that a) shrinkage should where possible be "light", and b) what "light" means needs to be determined by reference to estimation weights (rather than potentially dubious metrics such as retrospective bias). More generally, we should turn to models that use data (rather than *ad hoc* assumptions) to generate inferences.

XSA iteration convergence

XSA does not include a statistical estimation process in the usual sense, but rather uses an iterative estimation procedure that can be stopped before full convergence. The approach taken by ICES assessment working groups to the question of whether or not to converge varies widely. We show that the point at which the iteration is stopped can have a very significant affect on abundance estimates for a number of important ICES stocks. A comparison between an XSA run and an alternative exploratory state-space model for North Sea haddock shows that increased iterations also increases the *discrepancy* between the model estimates. We show further through simulation that there is a tendency for further iterations to move the assessment away from the underlying true population state. There are also indications that both the *q*-plateau age and the plus-group age appear to affect convergence, although this list of causal effects is by no means exhaustive. We conclude that a) it is essential to determine the convergence characteristics of any XSA assessments, and b) alternative methods need to be explored in cases where convergence is slow and leads to large changes in perceived stock dynamics.

State-space assessment models

Although there is (as yet) relatively limited experience and acceptance of state-space models in most ICES assessment working groups, they provide advantages over more traditional methods in a number of respects: a) they provide uncertainty estimates for stock metrics, b) they can accommodate observation error in catches, and c) they remove the need for *ad hoc* assumptions. They should be considered as valid alternatives in cases where these issues arise.

Survey-based assessment methods

We present work on two developments in the SURBA model. SURBA+ is an AD-ModelBuilder implementation that addresses several shortcomings in the original SURBA model: a) it models fishing mortality rather than total mortality, which is more useful for fishery managers but assumes a knowledge of natural mortality; b) it uses random effects approaches to smooth variations in mortality components, rather than *ad hoc* smoothing; c) it allows the age-effect in mortality to vary through the time-series, rather than being fixed as before; and d) it incorporates a recruitment model. We show the improvement in inference and management advice that these modifications can make for a sample case stock (3Ps cod). We also discuss briefly a parallel development in the original SURBA code, which is an implementation in the R package (SURBA-R). This may smooth the transition between the outdated current

SURBA code and the new SURBA+ code, and it is hoped that a single joint implementation can be developed in time.

Length-based assessment methods

We review recent work in length-based assessment methods, and collate conclusions on the utility of different approaches. This is a potentially valuable but also very difficult field that does not appear to have a natural home at the moment in ICES. We consider further an analysis of the sensitivity of a spurdog assessment to assumptions about early fishery selectivity for which there is few data, and find that the assessment is relatively robust to these assumptions.

Uncertainty in age–length keys (ALKs)

Through a simulation study, we demonstrate the effect of uncertainty in age–length keys on the assessment of roundnose grenadier in several Atlantic areas. We conclude that age-based assessments are unreliable for this stock because of ALK uncertainty, and suggest development of life-stage-structured approaches.

Future directions for WGMG

We suggest that the most useful way forward for WGMG in the short term could be a series of themed workshops for which WGMG would act as a steering group. The first of these could be a collation and comparison of assessment models from around the world, including many which are not currently used in ICES but which might bring benefits.

1 Introduction

1.1 Terms of Reference (ToRs)

The Working Group on Methods of Fish Stock Assessments [WGMG] (Chair: Coby L. Needle, UK) met in Nantes, France, from 20–29 October 2009 to:

a) Work according to specific ToRs developed intersessionally by the end of June 2009 in consultation with ACOM, relevant benchmark and assessment WG chairs, and relevant stock assessors. These ToRs are to be considered and finalized by SCICOM at the ASC meeting in September 2009.

b) Review the major problems and possible solutions to fish stocks assessments. The review should include an analysis of strengths and weaknesses, conditions for applicability of alternative solutions and process issues such as quality assurance protocols, sequential peer reviews and benchmarking.

c) Prioritize (in combination with ACOM) common methodological problems identified in benchmark reviews and recommendations by external reviewers.

The ToRs developed during summer of 2009, and subsequently agreed by SCICOM in September 2009, were as follows:

Number	Person	Group	Benchmark species	Request
1	Carmen Fernández	WGHMM	Hake	Development/application of assessment methodologies not reliant on age–length keys (Multifan-CL, Stock Synthesis, Gadget, global production models...)
2	Carmen Fernández	WGHMM	Hake	Accounting for revisions in growth parameters in assessments when there is no alternative way of ageing fish.
3	Carmen Fernández	WGHMM	Hake	Methods for reconstructing historical series of discards or alternative ways of coherently accounting for discards in assessments when there are many gaps in the series of estimates.
4	Mike Sissenwine	ACOM	General	Advice on appropriate shrinkage factors to be used for different stock situations, along with a review of shrinkage in other advisory areas (ICCAT, NAFO, IACCT, etc).
5	Coby Needle	WGNSSK	General	Simulation study of the relationship between convergence and population estimates for XSA (and other methods).
6	Harald Gjøsæter	WKSHORT	General	Discussion of case studies in which environmental information has been used in fishery assessment and management.
7	Lionel Pawlowski	WKDEEP	Roundnose grenadier	Evaluate the influence on stock assessment of uncertainty in age–length keys.

1.2 Report structure

Six Working Papers (WPs) were presented during the first two days of the meeting, and these are summarized in Section 2 with an abstract and discussion summary for each. Section 3 is in two parts: the first reviews developments in length-based assessment methods (ToR a1) to date, while the second reports the results of analyses of the sensitivity of a length-age spurdog assessment to assumptions about pre-1980 catch-at-age structure. XSA shrinkage (ToR a4) is discussed in Section 4, which covers the historical context of shrinkage and provides recommendations on how it should be used. XSA convergence (ToR a5) is approached in a similar fashion in Section 5. Section 6 contains a review of how environmental information has been used in assessments and advice (ToR a6), while Section 7 reports a sensitivity analysis on the effect of age–length key uncertainty on assessments for roundnose grenadier (ToR a7). Work on further testing of the state-space stock assessment model (SAM) is presented in Section 8. Section 9 then includes updates of developments on two implementations of survey-based stock assessment methods, while Section 10 gives conclusions on future directions for WGMG and a proposal for a themed Workshop next year.

Attendance at WGMG this year was insufficient to allow consideration of ToRs a2, a3, b and c. This issue is discussed further in Section 10.

Finally, the report provides sections on references, participants and recommendations, as well as a brief summary of the simulation model used to provide data for the XSA iteration convergence exercise (Section 5).

2 Working papers

The following table summarizes the working papers and presentations given at WGMG:

Number	Name	Title	ToR	Paper	Presentation
WP1	Benoit Mesnil	A history of shrinkage in ICES	a4	Yes	Yes
WP2	Anders Nielsen	Status of the state-space assessment model	a4, a5	No	Yes
WP3	José de Oliveira	Exploratory assessment model for Northeast Atlantic spurdog	a1	Yes	Yes
WP4	Coby Needle	XSA convergence	a5	No	Yes
WP5	Lionel Pawlowski	Overview of the Roundnose Grenadier stock assessment in ICES Vb, VI, VII and XIIb	a7	No	Yes
WP6	Noel Cadigan	Extensions to the SURBA model (i.e. SURBA+)	Extra	No	Yes

2.1 WP1 – Benoit Mesnil: A history of shrinkage in ICES

2.1.1 Abstract

Shrinkage is defined here as the use of the mean over a defined period or number of ages to modify (to a greater or lesser extent) population or fishing mortality estimates. It is generally employed in an attempt to reduce the affect of fluctuations at the end of a time-series. The WP retraces the views of WGMG on shrinkage, as related in its report since 1984. Shrinkage was first introduced in the context of predicting recruitment from multiple index series, as a mechanism to improve the precision of predictions. The justification to use calibration regression and shrinkage was provided by an important paper by a professor in statistics discussed at the 1984 WGMG meeting (ICES-WGMG 1985). The topic was again reviewed at the 1987 meeting (ICES-WGMG 1993a) where a paper by John Shepherd (which became the manual for RCRTINX) was discussed. However, WGMG disregarded the statistical foundation of

shrinkage and only drew attention on its practical aspect ("The key question is, in fact, to know whether or not it is useful to consider the past series, and especially its average value, as valuable first information."). WGMG returned on this in 1993 (ICES-WGMG 1993b) and basically endorsed the finding by Rosenberg *et al.* (1992) that calibration with shrinkage was the preferred method among the class of regression estimators.

In the context of VPA tuning, shrinkage was introduced later and in a more oblique way. The 1989 WGMG meeting set out to draw conclusions from the methods contest workshop in Reykjavik in 1988 (ICES-WGMG 1993a). Participants were impressed that time-series methods, which make use of signal in the recent period, performed much better than others. Moreover, the Laurec-Shepherd and Hybrid tuning methods used at the time were highly dependent on the quality of the terminal data points and often produced extravagant estimates of stock size and F, and thus very erratic TAC forecasts. Hence, the 1991 WGMG decided that some restraint on the variation of F estimates, as implemented in TSER, was needed (ICES-WGMG 1991). However, WGMG was reluctant to jump into the complexity of time-series methods and chose to use, by analogy, an *ad hoc* device familiar to most users: the shrinkage to the mean as in RCT. A more learned exposition of shrinkage is given in the 1993 report (ICES-WGMG 1993b), as a way to balance variance and bias. Trials indicated that a light shrinkage was beneficial in some instances, and this was reflected in the Blue Pages (ICES-ACFM 1995). However, WGMG points out that shrinkage produces systematic errors in the presence of a real trend in F and thus should be avoided (this was a clear message in Appendix 10 of Darby and Flatman, 1994). Moreover, WGMG notes that there is a theoretical explanation for why shrinkage reduces random variation in the predicted F, but the application of shrinkage to reduce retrospective patterns is still an *ad hoc* procedure without a satisfactory statistical basis; shrinkage can make the situation worse.

Overall, it is apparent that the implementation of shrinkage in VPA tuning is no more than an *ad hoc* device, to make the methods "idiot-proof" and reduce extreme variations in F, stock size and TAC estimates from one assessment to the next. The recommendations in previous reports were that users should explore the weight given to the mean, i.e. based on improvement in the retrospective pattern, despite the fact that no formal link has been established between shrinkage and the retrospective pattern. It is not judicious to use strong shrinkage as this is like confessing that the tuning data are merely worthless. In any case, it has always been made clear that shrinkage is inappropriate in cases where external information indicate that a trend in F is ongoing.

2.1.2 Summary of discussion

During the discussion following the presentation, it was highlighted that shrinkage in XSA should never be fully turned off. To produce abundance estimates for ages for which there are no data other than landings, XSA uses means which are derived through shrinkage. If shrinkage is simply absent, XSA will fill these ages with a predetermined dummy value that bears no relation to the stock in question.

The question of the use of shrinkage in non-European stock assessments was raised. The ADAPT method, prevalent in North America, does use a form of shrinkage, but only for estimating *F* on the oldest ages: which is a different issue from the population shrinkage implemented in XSA.

The meeting commented that we cannot *expect F* to conform necessarily to any kind of mean. This implies that the fishing industry would be able to choose deliberately to fish at a given mortality rate, which WGMG considers to be impossible. Fleets can remove a number of fish that they think will lead to a particular *F*, but in-year knowledge of stock abundance is not usually sufficient to allow this to be done with any accuracy.

The affect of shrinkage is not solely a function of the specified shrinkage SE, but will depend also on the number of surveys available and their variance. In other words, XSA shrinkage is not scale invariant, and effect of shrinkage needs to be tested each time it is used.

WGMG agreed that it would be useful to provide a practical demonstration of the problems with shrinkage highlighted during the presentation. Further work on shrinkage is provided in Section 4 below.

2.2 WP 2 - Anders Nielsen: Status of the state-space assessment model

2.2.1 Abstract

The state-space fish stock assessment model was summarized to the Group with a focus on the rationale behind using random effects to describe the underlying random variables that are not observed (fishing mortalities and stock sizes). Contrary to (semi-) deterministic approaches the state-space assessment model allows observation noise on observed catches, and is able to quantify those. Contrary to fully parameterized statistical assessment models, the state-space model has fewer model parameters, and the number of model parameters does not increase with every new year of data.

In addition the model has a number of appealing properties. It allows selectivity to gradually evolve during the data period, it allows missing data, and finally it estimates the underlying process noise, which is useful for forward predictions.

Previous implementations of state-space assessment models (Gudmundsson 1987, 1994; Fryer 2002) have been based on the extended Kalman filter, which uses a first-order Taylor approximation of the non-linear parts of the model. The current implementation is based on the Laplace approximation which is better suited to handle non-linearities, and further validated by importance sampling.

The state-space model has previously been validated by comparison to existing assessments and via simulated data. To further validate the model, it was extended to allow jumps in the underlying process to follow a mixture between a Gaussian and a fat-tailed Cauchy distribution, as opposed to a purely Gaussian. The model applied to North Sea Cod estimated the Cauchy fraction to be zero, and even forcing the Cauchy fraction to be 30% did not make the underlying process take noticeable sharper jumps.

It was demonstrated how the recent decision to change XSA shrinkage SE from 0.5 to 0.75 for Eastern Baltic Cod radically changed the perception of the stock in the final year to be more in line with the state space assessment model. The presenter argued against using *ad hoc* criteria for setting these shrinkage parameters.

A simple web interface (http://www.stockassessment.org) to the state-space assessment model was presented. Collaboration at assessment working groups is often reduced to one or two members doing the actual assessment modelling, and remaining working group members reviewing and commentating on the results only. Part of the reason most working group members don't even try to reproduce the assessment is that it takes a lot of work to get everything set up correctly. Typically several programs (specific versions) need to interact and the data need to be on a specific format. The web interface presented reduces this obstacle. Once the stock coordinator has set up an assessment all members can reproduce the assessment and all the resulting graphs and tables simply by logging in and pressing "run". The working group members can also experiment with the model configuration and input data and easily compare the results. It would clearly be beneficial to have more hands and eyes on the details of each assessment.

2.2.2 Summary of discussion

The web interface is currently set up for a number of specific cod stocks. It takes the presenter around 30 minutes to set up the process for a new stock – as yet there is no facility for stock assessors to do this themselves, although existing runs can be modified as required. The Group decided that it would be instructive to apply the method to data for North Sea haddock, to inform analyses on XSA convergence and shrinkage.

The Group commented that it is desirable to expand the number and range of output plots from the system, because the amount of time needed to do this in assessment Working Groups is substantial and detracts from other important work. Also, better diagnostic graphics help people better to understand what the model is doing.

Some Group members expressed concern over possible confounding between process and measurement error. Last year's simulation study was intended to address this, but the discussion indicated that a further demonstration would be beneficial.

The approach to shrinkage taken by the Baltic WG for the Eastern Baltic cod assessment was discussed. It is clear why they decided to use a shrinkage SE of 0.75 instead of 0.5, for that gave a stock estimate closer to that produced by the state-space model, but it was not clear why they decided to move from 0.5 in the first place. More generally, the approach taken to model settings and verification is not consistent across assessment WGs, and this needs to be addressed.

2.3 WP 3 - José de Oliveira: Exploratory assessment model for Northeast Atlantic spurdog

2.3.1 Abstract

An exploratory assessment model for Northeast Atlantic spurdog, developed for ICES-WGEF (2006), is presented. The model is based on an approach developed by Punt and Walker (1998) for school shark (*Galeorhinus galeus*) off southern Australia. It is essentially age- and sex-structured, but is based on processes that are length-based, such as maturity, pup-production, growth (in terms of weight) and gear selectivity, with a length-age relationship to define the conversion from length to age. Pup-production (recruitment) is closely linked to the numbers of mature females, but the model allows deviations from this relationship to be estimated (subject to a constraint on the amount of deviation). The model fits to a combined Scottish groundfish survey index of abundance, and to proportion-by-category data from both the survey and commercial catches (aggregated across gears). Four categories were considered for the survey proportion-by-category data, namely length-groups 16–31 cm (pups); 32–54 cm (juveniles); 55–69 cm (subadults); and 70+ cm (maturing and mature fish). The first two categories were combined for the commercial catch data to avoid zero values. The only estimable parameters considered are total virgin biomass (B_0), Scottish

survey selectivity-by-category (3 parameters), commercial selectivity-by-category for the two fleets (4 parameters two reflecting Scottish selectivity, and two England and Wales selectivity), and constrained recruitment deviations (1905–2005). The model assumes that there exist two commercial catch exploitation patterns that have remained constant since 1905, which is an oversimplification given the number of gears taking spurdog, and the change in the relative contribution of these gears in directed and mixed fisheries over time. This simplifying assumption allows stock dynamics to be taken back to near-virgin levels. The model estimates current depletion levels of around 5% relative to 1905, and 7% relative to 1955.

2.3.2 Summary of discussion

WGMG queried how the model could interpret a low catch in the 1920s as resulting from a high biomass and low *F*. This is largely based on an extrapolated selectivity pattern, which seems unfortunate as much of the current stock-state perception is driven by the estimates of high historical abundance. Such "heroic" assumptions are not unusual in studies that seek to reconstruct historical populations, but they do need to be justified. WGMG decided that an analysis of the sensitivity of the assessment to historical selectivity assumptions would be beneficial, and this is presented in Section 3 below.

2.4 WP 4 – Coby Needle: XSA convergence

2.4.1 Abstract

Before 2007, the assessment of haddock in the North Sea and Skagerrak was conducted by WGNSSK using the DOS version of XSA (Darby and Flatman 1994; ICES-WGNSSK 2006) in which the number of model iterations was truncated to 30. This was done for two reasons. Firstly, there was a perception that continuing to iterate XSA much beyond 30 would result (in certain situations) in a positive bias in stock abundance estimates; in other words, one possible response of XSA to noisy catchability residuals was to estimate a larger population, and this bias may increase with increasing iterations. Secondly, 30 iterations is the first point at which the user of the original DOS implementation was asked whether more iterations were required.

At the 2007 and subsequent WGNSSK meetings (ICES-WGNSSK 2007, 2008, 2009), the FLR version of XSA (FLXSA) was used the assess haddock. The default setting of this implementation is to iterate to convergence, and this is the approach now taken in all update assessments of that stock. The presentation given at WGMG demonstrated that repeating the 2007 assessment but with iterations truncated to 30 resulted in an estimate of SSB in 2006 that was around 60000 tonnes lower than the estimate from a fully converged assessment. Similar conclusions were reached for the corresponding North Sea whiting assessment, but not for North Sea cod. Simulation studies were rather inconclusive. Given that the difference in SSB estimates is roughly equivalent to the entire North Sea haddock quota for that period, understanding of the reasons for these results is important.

For a statistical catch-at-age assessment model, full convergence would be the only logical choice. However, XSA is an iterative procedure with unclear convergence properties, and it is not at all obvious that iterating to convergence results in an assessment that is closer to reality. The concern persists that further iterations may be inflating abundance estimates, as has been seen in the past.

2.4.2 Summary of discussion

WGMG expressed concern over this issue, and agreed that it was necessary to present the problem clearly. It was less clear what should be done about it. One suggestion was to run the state-space model on the stock, to determine what population dynamics would be estimated by a fully statistical catch-at-age model (both with and without variation in catches). Another suggestion was to try to develop two stock simulations, one with similar convergence properties to haddock, the other with no convergence, and try to determine what leads to the lack of convergence in XSA. The work carried out on this issue during this meeting of WGMG is detailed in Section 5 below.

2.5 WP 5 – Lionel Pawlowski: Overview of the Roundnose Grenadier stock assessment in ICES Vb, VI, VII and XIIb

2.5.1 Abstract

A review of the issues with the assessment of roundnose grenadier in ICES Subareas VI and VII, and Divisions Vb and XIIb was presented. Within ICES, the scientific basis for this stock identification is considered uncertain. This stock is generally considered to be in a data-poor situation. Therefore, only biennial advice is given with the recommendation that catches be constrained to 50% of the level of the beginning of the fishery (1990–1996), assuming no expansion of the fleets. Due to many sources of uncertainties, assessments using SVPA have been exploratory each year. This stock is scheduled for the benchmark process in 2010 and is also considered by the EU Deepfishman project which aims to review all available information on deep-water stocks.

Assessment methodology suffers from several problems. A frequent source of criticism is the use of SVPA considering only 19 years of data are available for a species with maturity at 8–14 years and a lifespan beyond 50 years.

Some uncertainties with landings statistics in XIIb exist which have led in recent years to the exclusion of landings from XIIb from the assessment. Discard data have been scarce and integrating the few data available requires making risky assumptions for the assessment. Age reading is known to require specific training. Few data are available which has constrained the use of an aggregating age–length key in the assessment despite substantial changes in the length distribution of the landings. Uncertainties in the ALK for this stock have been explored this year (ToR a7) by WGMG (see Section 7). There has also been an ICES workshop on age-reading (ICES-WKARRG 2007).

The biology and population structure of roundnose grenadier is also a challenge for the assessment as this fish occupies different depths according to its size, the larger individuals being in the shallower depths (500–750m), juveniles around 1000m, and intermediate sizes deeper (up to 1800m). Fishing effort at depth has changed through time, therefore the size structure of the landings reflects both the evolution of the fishery and the differences in length distribution at the various depth harvested by the fishing vessels.

Efforts on improving the assessment have been made during the recent ICES WGDEEP deep-water species working groups to quantify the effects of discards on the assessment, especially in the early days. Discards data are rather scarce and attempts have been made to extrapolate the few available datasets in order to rebuild catch for the 19 years of the whole time-series. Another approach combining fishing efforts at depth and the little information from scientific surveys on the vertical structure of the stock has been made to rebuild catch. All assessments suggest that the stock has strongly declined, and show consistent similar SSBs levels and fishing mortalities in recent years, the major differences between assessments being the estimates of biomass at the beginning of the 1990s. Results suggest however that integrating discards in recent years does not substantially change the results of the assessment (Pawlowski and Lorance 2009)

2.5.2 Summary of discussion

Much of the discussion was concerned with further information on the distribution, biological characteristics and fisheries of roundnose grenadier, so as to be better able to consider the best way forward for WGMG with regards to this stock. It is clear that data availability is poor, and this has implications for attempts to shoe-horn the few available data into a standard VPA-type assessment approach. As a first step, it was suggested that it would be helpful to run sensitivity analyses of the effects of ALK uncertainty on subsequent assessments. Results of these analyses are given in Section 7 below.

2.6 WP 6 – Noel Cadigan: Updates on Noel's version of SURBA

2.6.1 Abstract

We derive some basic statistics that describe the variability of a survey index derived from stratified random sampling for several Northwest Atlantic fish stocks. We also show how the survey variance component can be incorporated into stock assessment models like SURBA or ADAPT. In addition, we show how additional "non-survey" variability related to interannual changes in catchability can be incorporated into stock assessment models. Quasi-likelihood methods based on the means and variances of survey indices, relative to the stock as a whole, are used to develop an estimation procedure that incorporates survey sample sizes and estimates of within-survey variability. This may lead to improved estimation of stock size, in terms of more precise parameter estimates that are less sensitive to poorly sampled ages.

2.6.2 Summary of discussion

The presenter confirmed that the method described can only be used to estimate survey variance from a randomly stratified sampling design. There followed discussion on where this might be relevant in Europe, where most surveys are conducted using fixed stations. The Scottish monkfish survey was suggested as a potentially tractable example. Some of the discussion focused on details of the plots given in the presentation. There was also a comment that versions of kriging might achieve the same result in a more efficient way – the presenter replied that his approach was quite parameter-heavy, but also appeared relatively robust.

The presentation showed a method by which the age-pattern in survey catchability can be estimated along with stock abundance, even without commercial catch data. Survey catchability is usually fixed in survey-only models such as SURBA. This has clear implications for ongoing work with survey-based assessments and several further modifications were suggested – these are explored in Section 9.1 below.

3 Length-based assessment methods (ToR a1)

3.1 Developments on length-based assessment methods (ToR a1)

Within ICES, the Study Group on Age-Length Structured Assessment Models (SGASAM) aimed to address the issues concerned with the use of length structure into stock assessments methods. As other groups with the relevant expertise exist within ICES, the members of WGMG consider this ToR requires a dedicated workshop to review the current developments on length-based methods. The following section is however a short overview of the subject from last SGASAM report (ICES-SGASAM 2006) and a PhD. manuscript on the length structured modelling of the northern stock of European hake (Drouineau, 2008).

Overview

There are many stocks within the ICES area for which it is acknowledged that age-based assessments are inappropriate and where the use of length-based methods should be considered. Such situations occur when:

- length based models are considered to give a better representation of biological and fishery processes;
- age-based data are unreliable or unavailable compared with length-based data;
- age is not considered to be a good proxy for length.

Length-structured models have the advantage of allowing a good description of biology (such as predation or maturity) or harvesting (such as selectivity) without having to convert the length information into age (or *vice versa*). Using length information also implicitly takes account of interannual or interindividual variation. The interest in the length-structured approach has been growing in recent years with various levels of complexity and objectives. Length-structured models are commonly used to perform stock assessments, to estimate unknown parameters of the population dynamics, and to evaluate management plans or ecosystem models (Drouineau, 2008).

For stock assessment, the underlying population dynamics model generally only includes recruitment, growth and natural mortality. Additional processes may include migration and cannibalism. The fishery modules range from simple separable models (Kristensen *et al.*, 2006) to the representation of multiple fleets (Frøysa *et al.*, 2002).

Main models reviewed by SGASAM in 2006

- Stock Synthesis 2 ("SS2") is an assessment model (http://nft.nefsc.noaa.gov/test/SS2.html) which includes age and size-based population dynamics and observational phenomena such as ageing imprecision and is coded in ADMB (Dave Fournier, Otter Research Ltd.). Data include catch by fleet in weight or numbers, fishery and survey age and length composition, mean length-at-age, age composition conditional on length/gender, survey abundance, fishery cpue, mean body weight, and percentage discard by weight. The time-step is typically annual, but multiple seasons of varied duration can be defined. The population of each gender can be divided into a set of phenotypic morphs, each with unique growth and natural mortality parameters. Numbers-at-age for each morph are tracked independently, so that size-specific fishing mortality will have a differential effect on the survivorship of each morph. Expected values for data from each morph are accumulated within each gender to match the level at which observed data are collected. Growth parameters can be estimated internally to evaluate the effects of size-selectivity and ageing imprecision on observed length-at-age. Fishery age and length data can be specific to discard or retained samples, so provide necessary information to allow the model to estimate retention functions. Model parameters can be a function of environmental data or vary randomly or in time blocks. SS2 includes routines to estimate MSY and levels of exploitation that correspond to various standard fishery management targets. A user-selected harvest policy is used to conduct a forecast as part of the final phase in running the model. Parameter estimation occurs in a Bayesian context and the Monte Carlo Markov Chain algorithm is used to provide non-parametric confidence regions on parameters and derived quantities. In addition, SS2 is designed to produce a set of parametric bootstrap datasets. Comparable confidence regions on model parameters and derived quantities have been observed using the inverse Hessian, parameter profiles, MCMC, and re-running the model on the bootstrap data. In 2005, SS2 was used to assess the status of about 20 groundfish stocks off the west coast of the US

- LCS (WP1, ICES-SGASAM 2006) is currently under development at IMR, Norway and uses an approach similar to that used by Stock Synthesis 2 for incorporating growth. The method uses a Lagrangian approach where the population consists of a group of "super-individuals" each with its own growth characteristics and abundance which are projected forwards in time. The method has been applied to the Northern hake stock and also North Sea sprat.
- Multifan-CL (Fournier *et al.*, 1998), Fleksibest (Frøysa *et al.*, 2002) and A-scala (Maunder and Watters, 2003) can be used for stock assessment but also to estimate unknown parameters of the population dynamics such as growth and to some extent migrations. Multifan-CL can be spatialized.
- GADGET (*G*lobally applicable *A*rea *D*isaggregated *G*eneral *E*cosystem *T*oolbox) is a software tool (www.hafro.is/gadget) that can run complicated statistical ecosystem models, which take many features of the ecosystem into account. Gadget works by running an internal model based on many parameters, then comparing the data from the output of this model to "real" data to get a goodness-of-fit likelihood score. These parameters can then be adjusted, and the model re-run, until an optimum is found, which corresponds to the model with the lowest likelihood score. Gadget allows the inclusion of one or more species, each of which may be split into multiple stocks; multiple areas with migration between areas; predation between and within species; maturation; reproduction and recruitment; multiple commercial and survey fleets taking catches from the populations.

Modelling of the population structure

In length-based models, maturity and fecundity can be modelled through a maturity ogive depending on length (rather than age) or through a stochastic process where each length class has its own probability of maturity. The stock is generally divided between mature and immature individuals. GADGET allows making maturity dependant on a condition factor (Begley and Howell 2004)

Natural mortality is generally supposed to be known and constant in stock assessment models. For modelling efforts where trophic interactions are important or if cannibalism is strong (Frøysa *et al*., 2002), natural mortality has to be described differently. Cannibalism may be described as a function of the size of the prey and predator abundance. GADGET integrates a preference function depending on the respective sizes of predators and prey (Begley and Howell, 2004).

Growth is a key process for any length-structured model. In an ideal world, the model must be able to describe the average growth but also its interindividual variations (Chen *et al*., 2003). Several approaches exist to describe this process:

- It can be assumed that individual length within an age group follows a distribution around an average defined by a growth curve (Fournier *et al*., 1998, Maunder and Watters 2003). This requires some *a priori* knowledge of growth or at least some age-structured data. This approach does not allow the identification of the effect on fishing over the individual sizes of the stock.
- Another approach uses growth increments. This is probably the most used approach in length-structured models. Growth over a time-step follows a stochastic distribution from a growth curve to estimate the probability of moving to the next size class. This approach relies on a transition matrix (Sullivan *et al*., 1990, De Leo and Gatto 1995, Cruywagen 1997). This type of model however does not allow us to take account of genetic differences between individuals.
- A third method incorporates interindividual variation of growth through the assumption that parameters from a growth curve follow a particular statistical distribution (Sainsbury 1980, Smith *et al*., 1998, Smith and Botsford, 1998, Pilling *et al*., 2002). This idea is more adapted for Individual Based modelling and often used to estimate growth in tagging-catch and release programs (Laslett *et al*., 2002, Eveson *et al*., 2004). It can be very computationally intensive.

Data type used in length-based models

Catch data are often used for length-structured model calibration. This includes using length data (in weight and numbers) disaggregated or not by time and fleet and length class (Sullivan *et al*., 1990, Frøysa *et al*., 2002). Some assumption can be made that catch is a random variable where the average is predicted by a model. This is the case for Fleksibest. In other cases, total catch in weight and numbers are used separately from landings data collected at the fishmarket.

Fishing effort (when available) can be used as an abundance index and in that case is treated like if arising from scientific surveys (Frøysa *et al*., 2002, Breen *et al*., 2003, Punt 2003).

Abundance indices from surveys can be considered as following statistical distributions predicted by a model (Frøysa *et al*., 2002, Breen *et al*., 2003) or can be decomposed into a global abundance index with a probabilistic distribution of the composition of the indices per length-class (Fu and T J Quin II 2000, DeLong *et al*., 2001).

Tagging catch and release data are used to estimate growth parameters but sometimes also in length-structured models (Breen *et al*., 2003, Punt 2003).

Age-length structured models for the assessment of stocks

The last SGASAM report (ICES-SGASAM 2006) reviewed the use of these types of models when age data are sparse or unreliable. The stocks included were *Nephrops*, Northeast Atlantic spurdog, Northern hake and sprat, and Bay of Biscay hake. Details of the models used are available in the SGASAM reports (e.g. ICES-SGASAM 2006).

In Drouineau (2008), a length-structured and spatialized model of the Northern hake stock is detailed and fitted against datasets from the fishery. This model aims to estimate unknown parameters of the population dynamics and perform diagnostics of the stock. Growth and migration parameters are easy to set up but adjustment to observations is difficult probably because of the complexity of the model and the low quality of the available data. The mean growth-rate of the population for this stock has been estimated to0.124 y^{-1} which is lower than those estimated from scientific survey through tagging programs. Migrations appear to be well simulated and biomass estimates are close to those from XSA, although recruitment estimates are different. This model has not been evaluated by an ICES working group in the context of an assessment.

3.2 Sensitivity of spurdog assumptions about pre-1980 catch-at-age structure

WP3 describes the exploratory length-age model applied to the Northeast Atlantic spurdog stock, presented to ICES working group WGEF in 2006 (ICES-WGEF 2006), together with an addendum correcting some of the equations and suggesting extensions that are currently being pursued. When the working paper was presented to WGMG, concern was expressed about the projection of the model back to 1905, covering a large period for which only total landings data (expressed as tons landed) was available (more detailed data were only available from the early 1980s onwards). This backwards projection required assumptions to be made about how the fishery was split into different fleet components, and what the selectivity-at-age was for each of these components. The base run presented in WP3 (and shown here) assumed two fisheries, one with a Scottish selectivity (reflecting mostly a mixed demersal fishery) and one with England and Wales selectivity (reflecting mostly a longline and gillnet fishery), with the split in catches between the two being based on the average for the period 1980–1984. The work here therefore looks at sensitivity of model estimates to these assumptions.

Three alternative selectivity-at-age scenarios were considered and compared to the base run. The base run selectivity-at-age curves are shown in Figure 3.2.1(a) and two of the three alternative runs in Figure 3.2.1(b). The two alternative runs reflect a selection favouring older fish (Oldsel), and one favouring younger fish (Youngsel). These selectivity-at-age curves for the pre-1980 period (1905–1979) were derived by multiplying the estimated selectivity-at-age curves for the post-1980 period by the multipliers shown in Figure 3.2.2(a). The third alternative run (not shown in Figure 3.2.1) reflects full selectivity (Fullsel), with a selection of 0 for age 0 and 1 for all other ages for both fleets and sexes. Additional runs were also considered, assuming all pre-1980 selectivity reflected either post-1980 Scottish selectivity or post-1980 England and Wales selectivity, but these results yielded very little difference compared to the base run, so are not shown. Selection is actually length-based, so Figure 3.2.2(b) is given to show the conversion from length to age.

Results of the sensitivity analysis are shown in Figure 3.2.3. These appear to be relatively insensitive to the selectivity-at-age assumptions for the pre-1980 period. Table 3.2.1 also indicates that estimates of current depletion levels are also relatively insensitive to these assumptions, and range from 4.3 to 5.8 relative to 1905, and from 5.8 to 7.8 relative to 1955.

Table 3.2.1. Model estimates of current depletions levels (in terms of total biomass) relative to 1905 and 1955 (B_{depl05} and B_{depl55} respectively). CVs are shown in smaller font in square parentheses.

	BDEPL05	BDEPL55
Base run	5.1 [29%]	6.9 [28%]
Fullsel	4.9 [30%]	6.6 [29%]
Oldsel	5.8 [30%]	7.8 [29%]
Youngsel	4.3 [28%]	5.8 [27%]

In conclusion: although the extension of the model back to 1905 relies on strong (and hard to justify) assumptions about selectivity in years for which there are no data, a sensitivity analysis has demonstrated that model fits and conclusions are not sensitive to these assumptions. Therefore, the presence of such assumptions does not appear to invalidate the model.

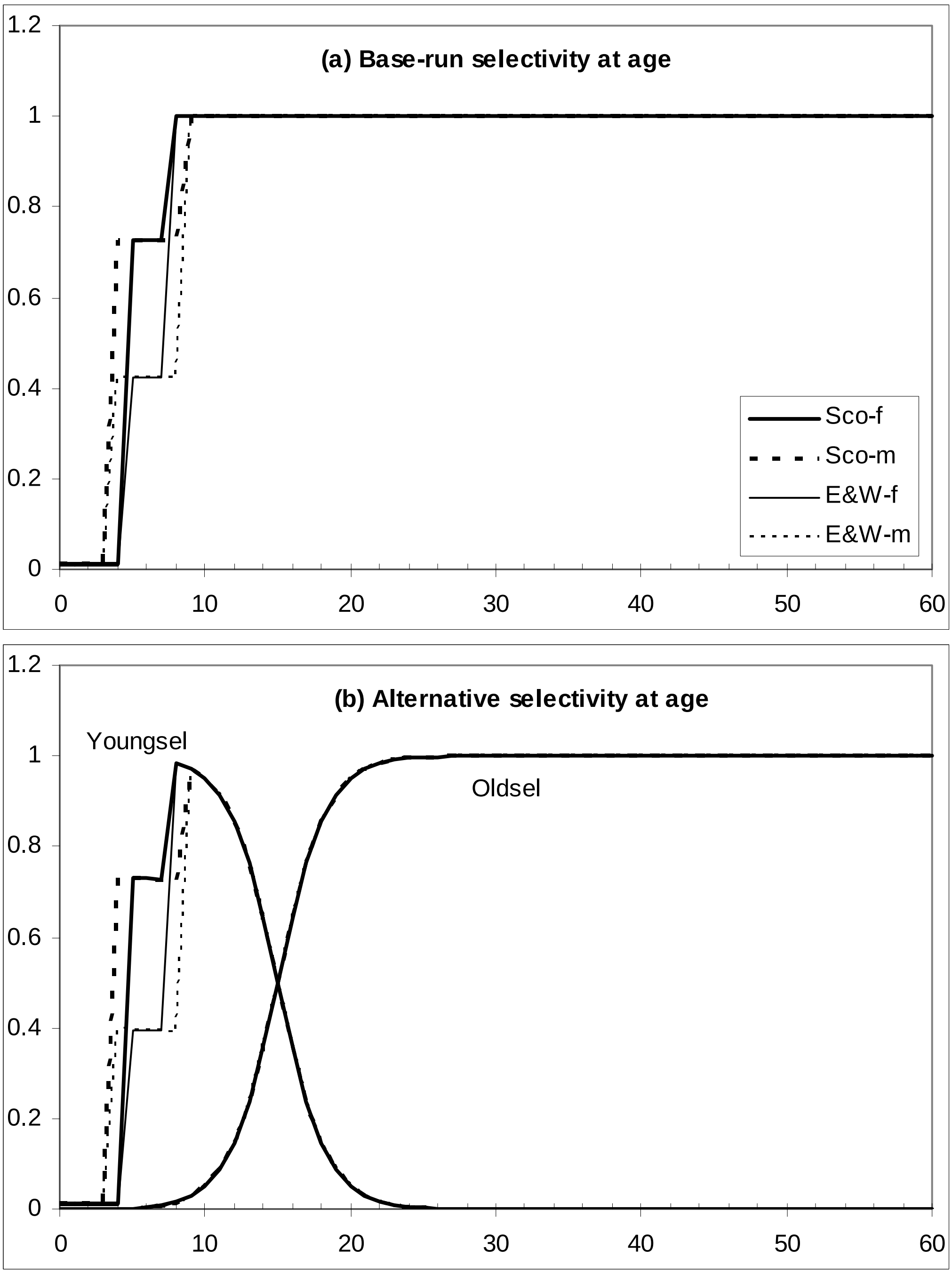

Figure 3.2.1. Selectivity-at-age curves for Northeast Atlantic spurdog, for two fleets (Scottish: Sco and England and Wales: EandW) and for both sexes (males: m, and females: f), with (a) reflecting the base-run as shown in WP3, and (b) reflecting two alternative runs where selectivity prior to 1980 favours young fish (Youngsel), and where it favours older fish (Oldsel). The alternative selectivity-at-age curves shown in (b) were derived by applying the multipliers given in Figure 3.2.2(a) to the selectivity-at-age curve for the post 1980 period in each case. [Note, the curves for both fleets and sexes fall on top of each other as demonstrated in the case of Oldsel in (b).]

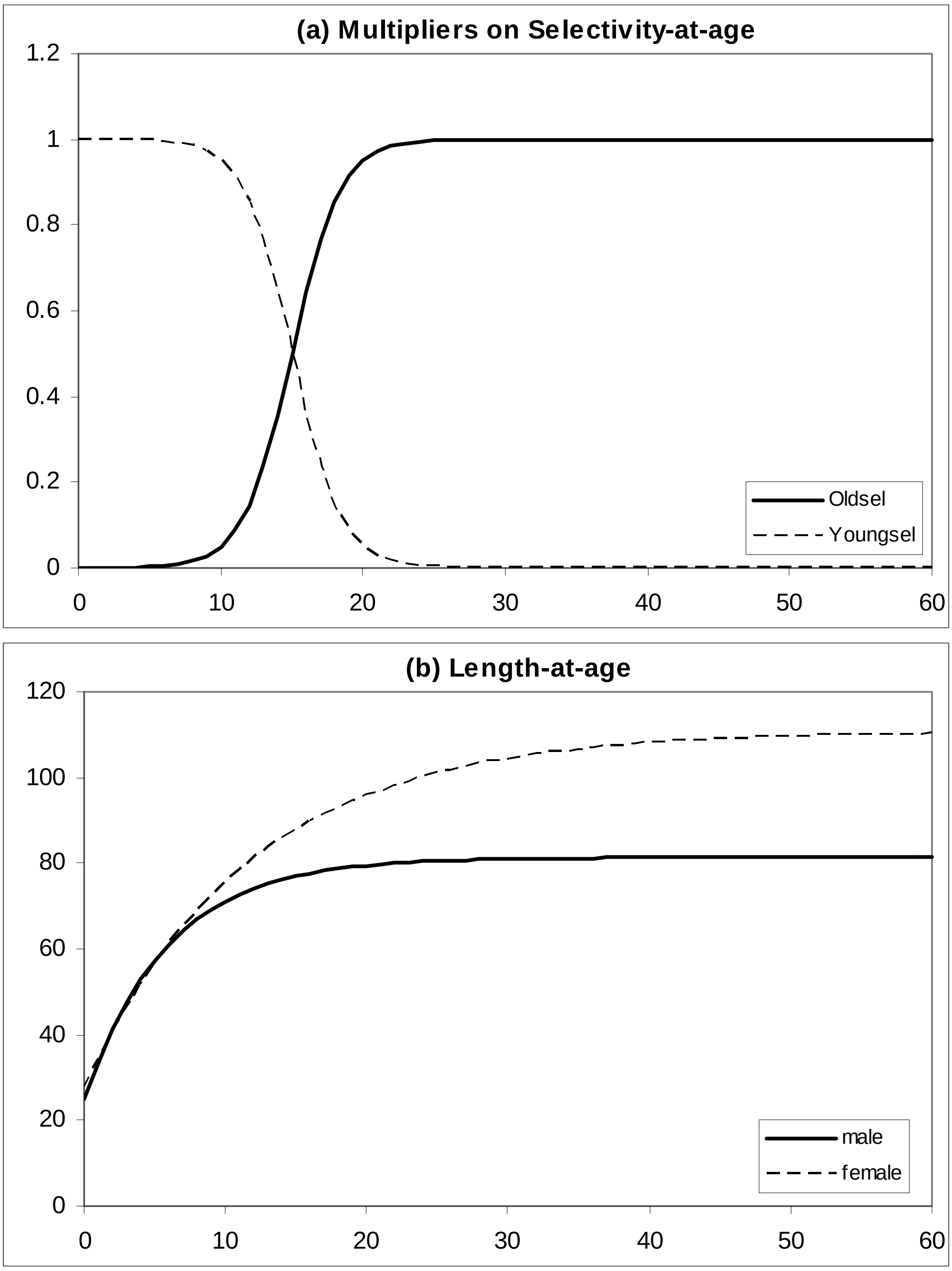

Figure 3.2.2. Additional information to help interpret results. (a) describes the multipliers that are applied to the post-1980 selectivity-at-age curves to derive the selectivity-at-age prior to 1980 for both fleets and sexes in the case of the two alternative runs reflecting selection favouring older fish (Oldsel) and that favouring younger fish (Youngsel). (b) shows the length-at-age curves, which is derived from growth curves based on length for each sex (see WP3).

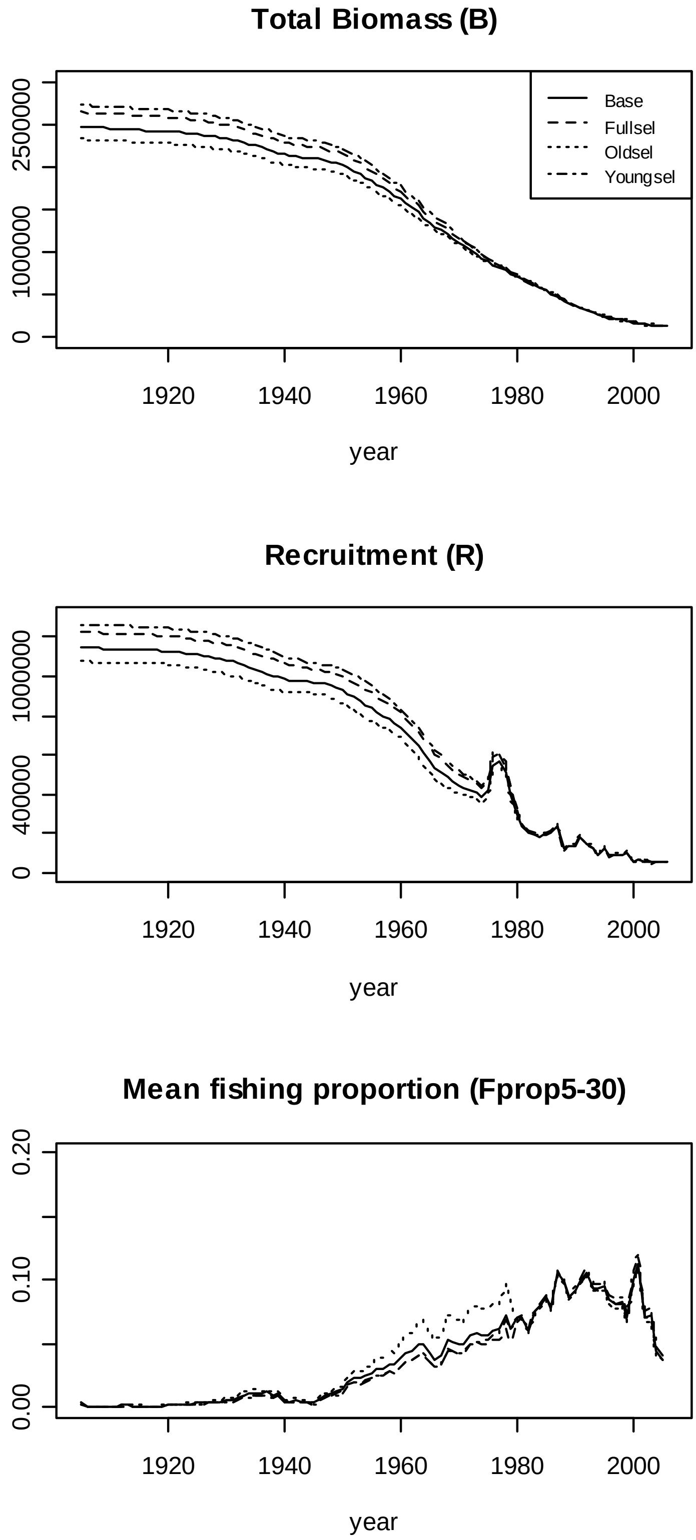

Figure 3.2.3. Sensitivity of estimated population trajectories (total biomass, recruitment and mean fishing proportion for ages 5 to 30) for alternative assumptions of selectivity-at-age shown in Figure 1 (Base, Oldsel, Youngsel). The full selectivity-at-age option (Fullsel, reflecting a selection of 0 at age 0 and 1 for all other ages) is not shown in Figure 3.2.1.

4 XSA shrinkage (ToR a4)

In the standard implementation of the VPA suite, three forms of shrinkage are available: i) shrinkage to the population mean, only applying to the so-called ages 'treated as recruits' where a non-linear relationship between index and population size is allowed; ii) shrinkage to population numbers-at-age, derived from an average F in recent years and catches in the final year, in the estimation of survivors at age in the final year ('year-shrinkage'); iii) shrinkage to mean F over some earlier ages used to estimate starting numbers at the oldest true age in each year ('age-shrinkage'). Item i) is not considered further here. Although ii) and iii) cannot be disconnected in the current ICES suite, they have different implications and will be dealt with separately.

4.1 Year-shrinkage

Year-shrinkage has an affect on estimates of stock size in the final year and hence on the TAC advice. Concerns with misuse of this option have often been voiced in ICES, and may be a reason why this term of reference is again addressed to WGMG (as in 1993).

In essence, the idea is (was) that if one intends to predict the current F (say) and if no major changes in effort, capacity etc. are known to have taken place recently (an assumption of no trend), then the mean of recent estimates of F is a sensible starting value, potentially associated with low variance. That is, when tuning involves relatively noisy survey or cpue indices, it may be of interest, in order to reduce the mean squared error of the predicted F, to combine the high variance estimates based on indices with the low variance estimate based on the mean, despite the bias carried by the latter (ICES-WGMG 1993b). To allow the weight given to shrinkage to vary depending on the quality of information, a weighted average procedure is used in which the weights are the inverse of the variance of each estimate. For the shrinkage mean, the user specifies a CV which is functionally equivalent to the SE of the estimated log-q's (Darby and Flatman 1994). The *a priori* shrinkage weight is then $1/CV^2$; hence a high CV implies a weak shrinkage (and vice versa).

As recalled by WP1 (Section 2.1), shrinkage in VPA tuning was introduced around 1991 to reduce the wide fluctuations in assessments and advice produced by the methods of the time (Laurec-Shepherd and Hybrid). These computed a catchability for each fleet then inferred the population size or F in the terminal year based on the cpue or survey index available for that final year only. Noise in that single data point was carried straight into the stock estimates causing embarrassing revisions of assessments and advice from year to year. A device was needed to restrain such fluctuations but proper approaches based on time-series methodology were deemed too complex for lay users. Hence, ICES resorted to shrinkage, a device already implemented in the recruitment prediction routine making it familiar to most people.

The recommendation in the Blue Pages (ICES-ACFM 1995) was: "*A low level of shrinkage (S.E. = 0.5) is suggested as a starting point in the VPA tuning. This level of shrinkage has been found beneficial in most cases. It is advisable, however, to explore other S.E. values using retrospective analysis. The number of years used in the shrinkage is normally five. If there are clear indications of a change in F within the last 5 years use fewer years.*"

Digging through the WGMG reports since the early 1990s, it is quite clear that shrinkage was viewed, even by its proponents, as no more than an *ad hoc* device with fragile theoretical bases. The experience is that it has often improved the stability of results from one assessment to the other, but the reduction in variance is by no means guaranteed (ICES-WGMG 1993b, p.12). WGMG also cautioned that the formal mechanism whereby shrinkage might improve retrospective patterns remains elusive; the 1993 meeting even warned that shrinkage could make things worse, notably when bias occurs in the converged part of VPA. In this respect, the recommendation to use retrospective analyses to adjust the shrinkage CV is odd.

WGMG has also made the message clear that shrinkage is inappropriate in cases where there is a trend (up or down) in effort as, for obvious reasons, it delays the ability of the assessment to detect or track the change in F. It is also inappropriate to recruitment fisheries, where large interannual variations in F may occur if management does not adjust appropriately to TAC advice, or if the size of recruiting year-classes are uncertain. However, it can be sensible to consider some shrinkage when one suspects a recent index has problems (i.e. acute year effects or negative Z's).

Another difficulty identified by the current meeting of WGMG is that, for a given CV, the actual weight given to the shrinkage estimate is context-dependent; it can vary with the number of index series, their relative precision, and it changes with the age considered. Only a close examination of the XSA diagnostics for each age/year-class, in the column named 'Scaled Weights', can allow one to realize what the exact affect of shrinkage is for that age. The default value of 0.5 in the software can mean a light shrinkage (as suggested by the Blue Pages) in some cases, but a strong one in others (e.g. when the signal from the tuning indices is weak).

The question then remains about whether shrinkage should be used at all, given its *ad hoc* nature. A number of sound approaches exist to fulfil its initial intention that is to allow for drifts in F over time while preventing sudden jumps just caused by noise in the data. For example, the state space approach (SAM) discussed in Section 8 precisely does this, in a respectable manner. Software packages to compute integrals of high dimensional likelihood functions are now easily available, and the past reservations against time-series or state space model formulations have no reason to persist. If an XSA run is still needed, diagnostics from runs with such methods should indicate whether a light or medium shrinkage is appropriate.

In any case, it is not judicious to accept the idea that a strong shrinkage is required. Reviewers will immediately interpret this to mean that the tuning data (and perhaps other inputs) are worthless.

Year-shrinkage is optional in ICA (Patterson and Melvin 1996); if switched on, the results of the shrunk and 'normal' run are kept in distinct files for inspection. To the group's knowledge no other assessment package outside ICES implements an equivalent apparatus to year-shrinkage.

4.2 Age-shrinkage

In this case, the procedure is to combine index based estimates of F for the older true age with an average F for a range of younger ages. In effect, this forces the exploitation pattern to be relatively flat at older ages (or at least avoids unrealistic sharp bends). The purpose here is equivalent to the specification of a terminal selection (relative to 1 at the reference age) in separable models, or to setting the terminal F to be some fraction of F at some younger age as in some implementation of ADAPT; in that sense, these other models also implement a form of shrinkage.

Although the effect of age-shrinkage on terminal population estimates, and hence on advice, is less spectacular than that of year-shrinkage, it is far from neutral especially for stocks where fishing mortality is low and thus VPA convergence is slow. In a different context, tests on simulated data have shown that misspecification of the exploitation pattern could lead to significant errors, notably for separable models (NRC 1998).

Whenever catch data exist for years prior to the first year with tuning data, a minimal degree of age-shrinkage must be used in the current version of XSA to initiate the VPA at the older age using an average of earlier ages' F (otherwise, all past cohorts start from an arbitrary F of 0.65, strangely enough). If only this effect is desired, one may enter a very high CV (e.g. > 2) to inhibit any other effect of shrinkage for the recent period.

4.3 Recommendations

Alternative modelling approaches are now available to serve the same purpose as shrinkage, but with sounder foundation than the *ad hoc* device implemented in XSA.

If the concern is with time-series property of the data, then benchmark groups should consider a proper time-series methodology to adjust or estimate the weight given to shrinkage. However, even with such methods, one should check that the parameters are well estimated. In some situations, time-series models and estimation procedures have been shown to produce seriously biased parameters estimates (e.g. de Valpine and Hilborn, 2005). Clearly, the present procedures of *ad hoc* choices for the amount of shrinkage in ICES groups are ineffective to set the shrinkage weights because they are entirely context dependent (Figure 4.3.1).

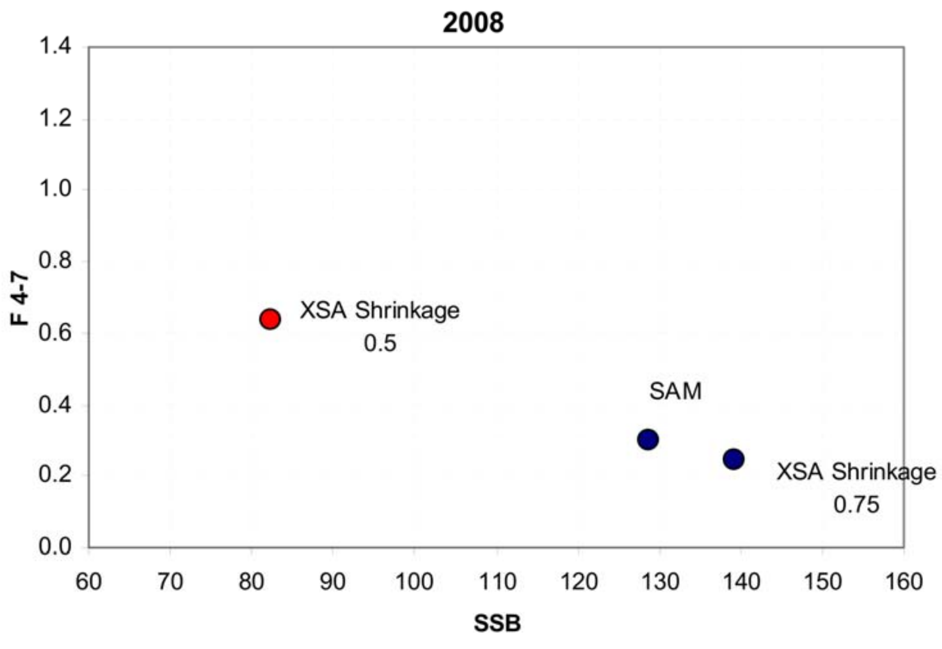

Figure 4.3.1. Spawning stock biomass and average fishing mortality for the Eastern Baltic cod stock in 2008 with two different shrinkage settings and with a state-space assessment model.

Determining the amount of shrinkage by minimizing retrospective patterns is not recommended because it could lead to seriously biased estimates of stock size or their trends.

Generally, only light shrinkage should be used to avoid introducing bias (consistent with WGMG recommendations since 1991). 'Light' should be measured by the actual scaled weights tabulated in the XSA diagnostics. If shrinkage has a large effect on assessment results, this is indicative that the wrong model is being considered.

Although they serve a different purpose, year- and age-shrinkage are switched on in the same menu in XSA and a single CV is applied in the standard VPA suite. A version where the two are disconnected is available (C. Darby, CEFAS) and should if possible be implemented in ICES.

5 XSA iteration convergence (ToR a5)

5.1 Introducion

This Section addresses ToR a5, which was proposed by Coby Needle (Scotland) and arose from concerns expressed by ICES-WGNSSK (2009):

> *Simulation study of the relationship between convergence and population estimates for XSA (and other methods).*

The background to this request is the recent history of the assessment for North Sea haddock carried out by WGNSSK, as summarized in Section 2.4 above. The analyses described below continue the work presented by Needle (WP5), and attempt to address the following questions in a generic sense:

1) Should XSA runs be iterated to numerical convergence in all cases?

2) Are assessment Working Groups applying consistent convergence criteria?

3) What XSA run settings are likely to affect convergence?

5.2 Previous advice on XSA convergence

The ICES Blue Pages, written in 1995 as a user manual for ICES stock assessment working groups, contain very little about the issue of convergence, and merely note (in a section on examples of XSA diagnostics) that:

> "The tuning has not converged after 40 iterations. In this case the user has chosen to stop the XSA. The user might have continued with further iterations in steps of 10 iterations at a time. The differences between F in the last and the second last iterations are given for the last year." (ICES-ACFM 1995)

This cursory note says nothing about *whether* the user should have stopped the XSA. The Lowestoft VPA manual (Darby and Flatman 1994) is rather more informative. From pp. 30–31:

> "With some datasets the program may not reach a converged solution before generating extremely low (zero) values of F. This usually requires a large number of iterations (> 30). If this occurs the program may fail when calculating subsequent outputs. It is recommended that when using *ad hoc* tuning, the user monitors the residuals displayed after each set of iterations and does not progress beyond 30 iterations before stopping the tuning run and examining the diagnostics file. If convergence has not occurred, the F-at-age values for the final year, calculated during the final two iterations, are recorded and can be compared. They can be used to identify the ages which are not converging." (Darby and Flatman 1994)

Note that this text is in the section on *ad hoc* tuning, but it is to this text that the reader is referred when looking for details on XSA iteration convergence so the convergence procedure is likely to be the same.

It is clear from Darby and Flatman (1994) that the advice was to stop the iterations after 30 steps *then check* convergence criteria from the diagnostics file. There was no general recommendation to stop the process altogether after 30 iterations: rather, the user was to take advantage of the break provided to ensure that the algorithm was not spuriously generating extremely low values of F, as was thought to be occasionally possible.

During a series of ICES Workshop Courses on Stock Assessment, Chris Darby offered the following advice on XSA iterations and convergence (ICES-WKCFAT 2002):

- "During trial runs, it is best to stop at 30 iterations to examine ages that may be causing problems.
- Raising the age at which catchability is held constant introduces more parameters to the model and may increase number of iterations required for convergence.
- Too many iterations can be caused by errors in parameter selections.
- Large numbers of iterations can mean no solution to assessment.
- Check for F values decreasing with iteration count, may be heading for zero F."

Again, the advice to stop at 30 iterations is only intended as an exploratory step, to ensure that XSA is converging correctly.

Finally, in his paper on the XSA model, Shepherd (1999) offered the following:

> "The iteration is repeated until the maximum change of any estimated fishing mortality is less than some small value (typically 0.0001) which generally requires fewer than 100 iterations. Because the computation is very quick, no attempt has been made to accelerate convergence. It is in principle possible for a two-phase iterative process such as this to ''hunt'', alternating between high and low estimates, or even diverge, but no such behaviour has been observed when using the algorithm described here, based on the logarithms of survivors as the working variables, despite extensive use in both simulation tests and practical applications." (Shepherd 1999)

In other words, at the time of writing, non-convergence in XSA was not thought to be a problem. Furthermore, Shepherd did not mention the possibility of spurious generation of low F at high iterations.

It is interesting how earlier advice on the desirability of stopping XSA after 30 iterations in order to check for convergence evolved over time within some Working Groups into a general perception that XSA runs should not be continued at all beyond 30 iterations. This is clearly not what was intended by the original advice, but certainly became the *de facto* approach for a number of Working Groups using the Lowestoft VPA suite to run XSA. The increased use of FLXSA (FLR Team 2005) has reversed this trend, with relevant Working Groups now generally iterating XSA to convergence no matter how many iterations that takes.

5.3 ICES Working Group approaches to XSA convergence

WGMG decided that it would be instructive to conduct a straw poll of ICES stock assessments carried out so far in 2009, to determine a) which assessment methods were being used, and b) if XSA or FLXSA (the FLR implementation; FLR Team 2005) was used, whether the algorithm was continued to convergence and how many iterations said convergence required. Reports were covered from the 2009 meetings of

AFWG, NWWG, WGBFAS, WGCSE, WGDEEP, WGHMM, WGNSSK, and WGWIDE. The models used to provide final assessments in these reports are summarized in Table 5.3.1.

Table 5.3.1. Final assessment methods used for stocks considered in 2009 by AFWG, NWWG, WGBFAS, WGCSE, WGDEEP, WGHMM, WGNSSK, and WGWIDE.

Final assessment method	Total
No assessment	48
XSA	26
TV survey	9
FLXSA	7
Commercial data trends	6
Survey and landings trends	5
Gadget	3
Survey trends	3
Yield-pre-recruit	3
ASPIC	2
SAM	2
ADCAM	2
TSA	2
FLICA	2
SXSA	2
B-ADAPT	2
SURBA 2.1	1
Absolute abundance from surveys	1
ADAPT-type	1
NFT-ADAPT	1
Bespoke catch-survey model	1
Bayesian production model	1
SURBA 3.0	1
Separable VPA	1
SAD	1
ASAP	1
TASACS VPA	1
SMS	1
Bayesian catch-at-age model	1
Grand Total	137

Of 137 stocks, assessments (whether analytical or trend-based) were provided for 89. Of these, 33 (37%) were assessed using XSA or FLXSA. Table 5.3.2 shows how these different assessments approached the question of XSA convergence.

Table 5.3.2. XSA iterations for the 33 XSA and FLXSA assessments carried out by ICES assessment Working Groups during 2009. FLXSA output diagnostics do not indicate how many iterations were required for convergence (the iterations for North Sea haddock are given here because the code for that assessment was available to WGMG).

Working Group	Year	Species	Unit	XSA or FLXSA	Iterations to convergence
WGNSSK	2009	Plaice	VIId	FLXSA	Not specified
WGNSSK	2009	Plaice	IV	FLXSA	Not specified
WGNSSK	2009	Sole	VIId	XSA	72
WGNSSK	2009	Sole	IV	FLXSA	Not specified
WGNSSK	2009	Saithe	IV VI IIIa	FLXSA	Not specified
WGNSSK	2009	Whiting	IV VIId	FLXSA	Not specified
WGNSSK	2009	Haddock	IV IIIa	FLXSA	122
WGCSE	2009	Haddock	Vib	XSA	27
WGCSE	2009	Sole	VIIa	XSA	Not converged after 30
WGCSE	2009	Haddock	VIIb-k	XSA	27
WGCSE	2009	Plaice	VIIf,g	XSA	Not converged after 30
WGCSE	2009	Sole	VIIf,g	XSA	47
WGCSE	2009	Whiting	VIIe-k	XSA	25
WGCSE	2009	Plaice	VIIe	XSA	Not converged after 30
AFWG	2009	Cod	I II (NE Arctic)	XSA	Not converged after 30
AFWG	2009	Haddock	I II (NE Arctic)	FLXSA	Not specified
AFWG	2009	Saithe	I II (NE Arctic)	XSA	88
AFWG	2009	Greenland halib	I II (NE Arctic)	XSA	50
NWWG	2009	Cod	Faroe Plateau	XSA	34
NWWG	2009	Haddock	Faroe	XSA	42
NWWG	2009	Saithe	Faroe	XSA	26
WGBFAS	2009	Cod	Baltic 25-32	XSA	39
WGBFAS	2009	Sole	IIIa	XSA	71
WGBFAS	2009	Herring	Baltic 25-27, 28.	XSA	Not converged after 50
WGBFAS	2009	Herring	Gulf of Riga	XSA	35
WGBFAS	2009	Herring	Baltic 30	XSA	Not converged after 30
WGBFAS	2009	Herring	Baltic 31	XSA	No information
WGBFAS	2009	Sprat	Baltic 22-32	XSA	Not converged after 30
WGDEEP	2009	Greater silver sr	I, II, IIIa, IV, Vb,	XSA	Not converged after 50
WGHMM	2009	Hake	Northern	XSA	58
WGHMM	2009	Sole	Bay of Biscay	XSA	Not converged after 30
WGHMM	2009	Megrim (L. whiff	VIIIc and IXa	XSA	Not converged after 200
WGHMM	2009	Megrim (L. bosc	VIIIc and IXa	XSA	Not converged after 40

The way in which XSA convergence is treated differs widely between assessments. 26 of the 33 assessments provided information on convergence: those which didn't were mostly FLXSA runs, for which convergence statistics are not provided in the standard output. Of these 26 assessments, 11 were based on an XSA that had not converged: seven were stopped at 30 iterations, one at 40, two at 50, and one at 200.

5.4 Convergence tests with ICES Working Group data

Tests of XSA convergence were carried out during the current WGMG meeting for six real datasets: North Sea cod, plaice and whiting, and plaice in Division IIIa, from the 2006 WGNSSK meeting; North Sea whiting from the 2008 WGNSSK meeting; and North Sea haddock from the 2009 WGNSSK meeting. Ten runs were produced for each stock: the full time-series run, plus nine retrospective runs. Analyses were also carried out of the sensitivity of the runs to changes in the user-defined XSA q-plateau. Finally, single-fleet runs for each of the datasets were conducted, but these did not lead to detectable differences from full-fleet runs and are not covered further here.

As an example, Figure 5.4.1 gives the full time-series run for North Sea haddock from the 2009 WGNSSK meeting, which is the case which motivated this study in the first place. The top-right plot shows that the SSB estimate for 2008 increases from just under 180 kt at 30 iterations to just under 220 kt at convergence (122 iterations, which indicates very slow convergence). Mean F has a corresponding decline with increasing iterations. The top-middle and top-right plots demonstrate that the variation in

SSB is restricted to the last few years of the time-series. This plot cannot show, of course, whether iterating XSA to convergence *improves* the assessment or not.

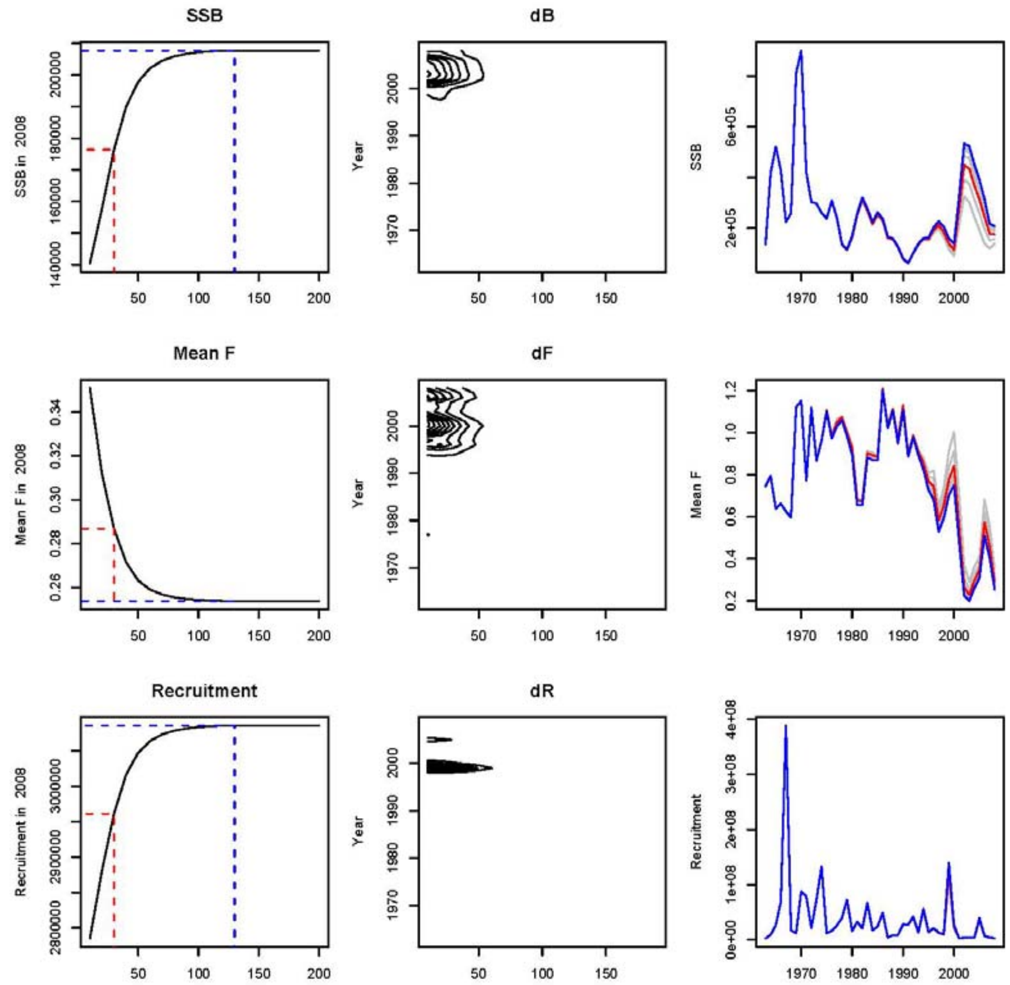

Figure 5.4.1. XSA convergence tests for North Sea haddock from the WGNSSK 2009 assessment. Top row: SSB. Middle row: mean F(2–4). Bottom row: recruitment-at-age 0. Left column: relationship between the estimate for 2008 and the number of iterations run (red lines indicate 30 iterations; blue lines indicate iterations required for convergence). Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line) and converged iterations (blue line) highlighted.

Another example is given in Figure 5.4.2, for North Sea whiting from the 2006 WGNSSK meeting. This plot uses the same XSA settings as were used in the WGNSSK meeting, and shows fairly rapid convergence (although with a strange point of inflexion in the convergence curve). This compares well with Figure 5.4.3, however, which is for the same stock with the q-plateau age increased from 5 to 7, with the result that the XSA no longer converges at all. This example supports the advice given by Chris Darby in 2002 (see Section 5.2), namely: "Raising the age at which catchability is held constant introduces more parameters to the model and may increase number of iterations required for convergence." It shows how easy it can be to generate an XSA assessment that does not converge, and how important it is for the convergence properties of these assessments to be checked as a matter of routine.

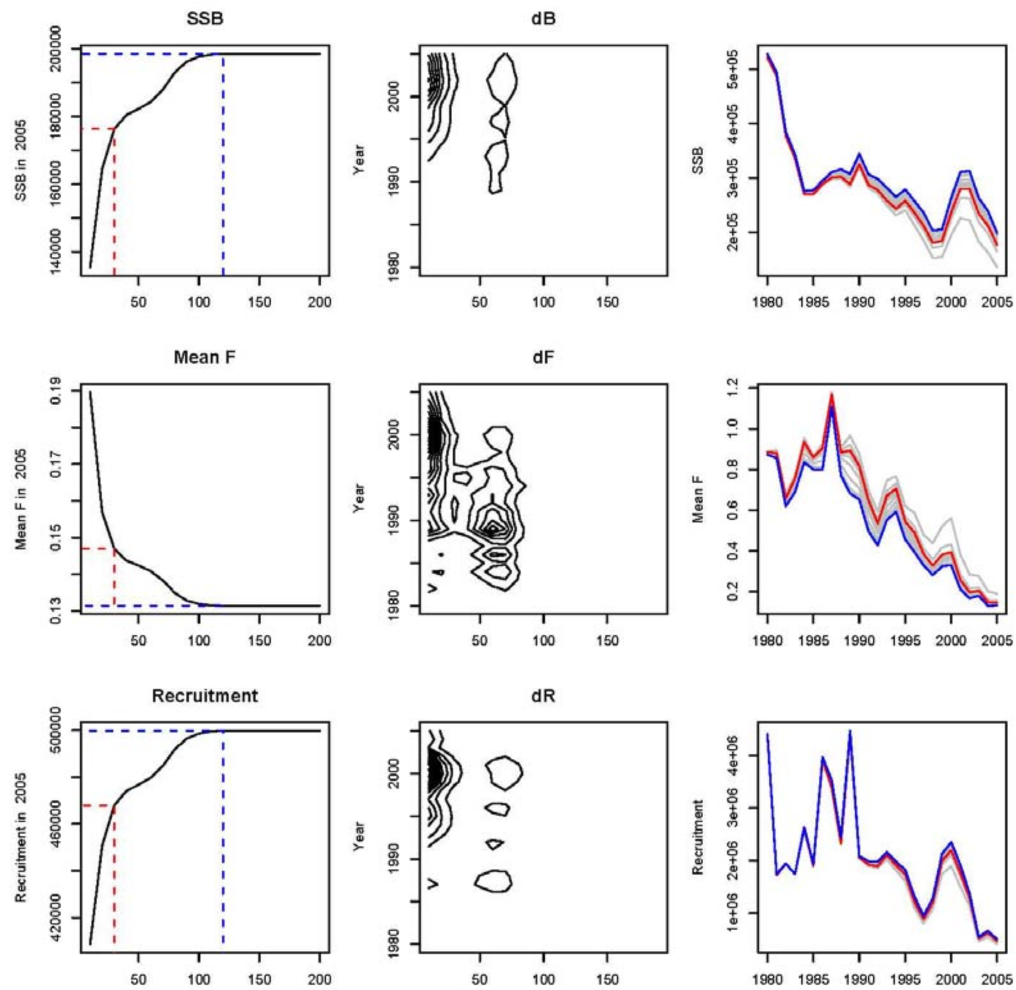

Figure 5.4.2. XSA convergence tests for North Sea whiting from the WGNSSK 2006 assessment. Top row: SSB. Middle row: mean F(2–5). Bottom row: recruitment-at-age 1. Left column: relationship between the estimate for 2005 and the number of iterations run (red lines indicate 30 iterations; blue lines indicate iterations required for convergence). Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line) and converged iterations (blue line) highlighted.

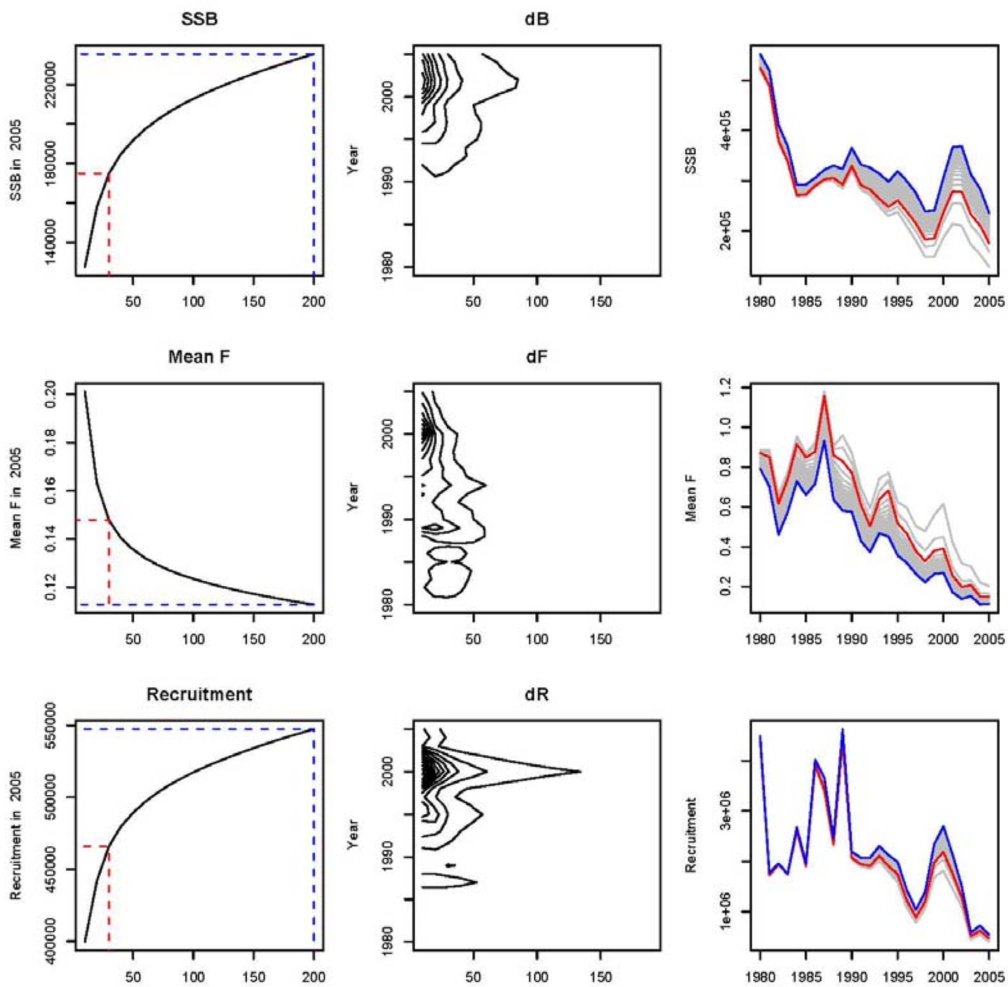

Figure 5.4.3. XSA convergence tests for North Sea whiting from the WGNSSK 2006 assessment, with the q-plateau age increased from 5 to 7. Top row: SSB. Middle row: mean F(2–5). Bottom row: recruitment-at-age 1. Left column: relationship between the estimate for 2005 and the number of iterations run (red lines indicate 30 iterations; blue lines indicate iterations required for convergence). Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line) and converged iterations (blue line) highlighted.

Many assessment scientists will recall experiences of XSA runs that did not converge. For example, the following note is taken from the report of the WGWIDE meeting in 2008, regarding southern horse mackerel: "The AMCI approach required strong conditioning and gave unrealistic results. XSA was used in 2006 and *did not converge*" [emphasis added]. An unconverged XSA run indicates that something is wrong: either the model is being used incorrectly, or the model is not appropriate to characterizing the available data. However, as we have seen from Table 5.3.2, there are many instances in current Working Groups of non-converged XSA runs being accepted as the basis for advice. The question of whether convergence actually leads to a better estimate is considered in the next Section.

5.5 XSA convergence for North Sea haddock: comparing XSA and SAM runs

5.5.1 Introduction

The motivating example for the ToR addressed in Section 5 is North Sea haddock. We have seen that the XSA assessment of this stock requires a large number of iterations to converge, and that the iteration process changes the final estimates of SSB and mean F considerably. WGMG decided to use the state-space assessment model (SAM) to generate an exploratory alternative assessment of the stock. The state-space assessment model is a statistical model and the estimates are based on maximum likelihood estimation, which has a very different way of diagnosing convergence. The state-space model is briefly described in Section 8 of this report. The XSA runs used for this comparative analysis are the same as those summarized in Figure 5.4.1.

5.5.2 Results

All runs of XSA and SAM showed the same overall trends in stock sizes and fishing mortalities (Figures 5.5.1 and 5.5.2).

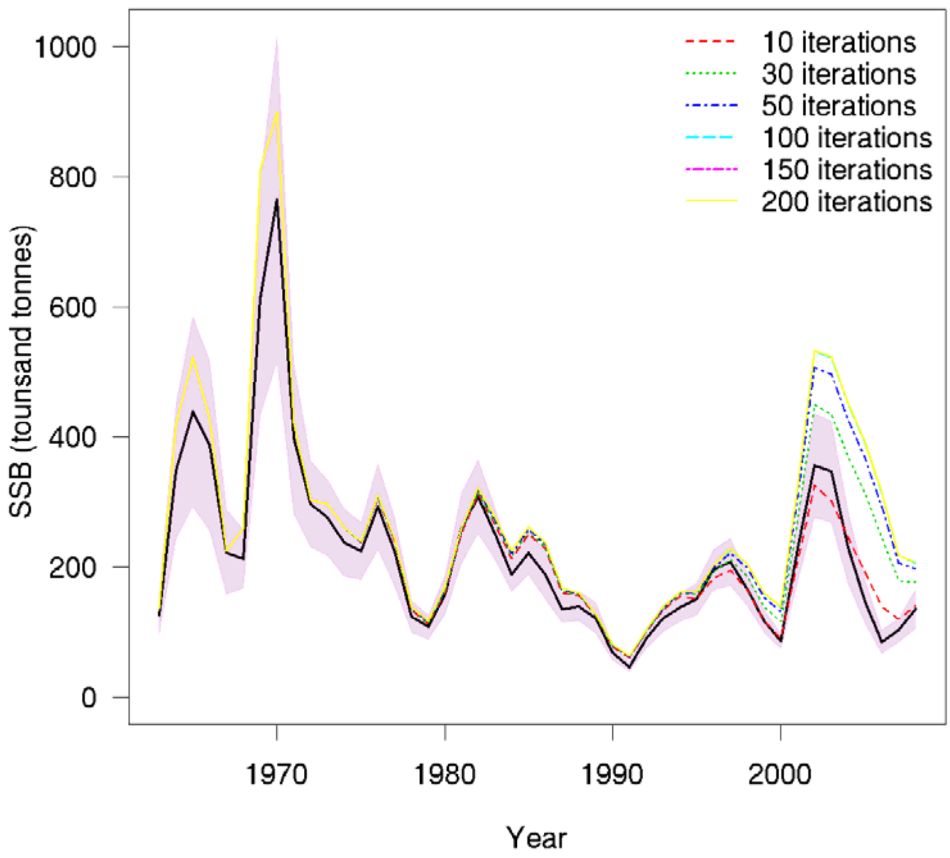

Figure 5.5.1. Spawning stock biomass estimated via the standard state-space model (thick black line) and corresponding 95% confidence interval (shaded areas), and by XSA with different numbers of iterations (dashed lines).

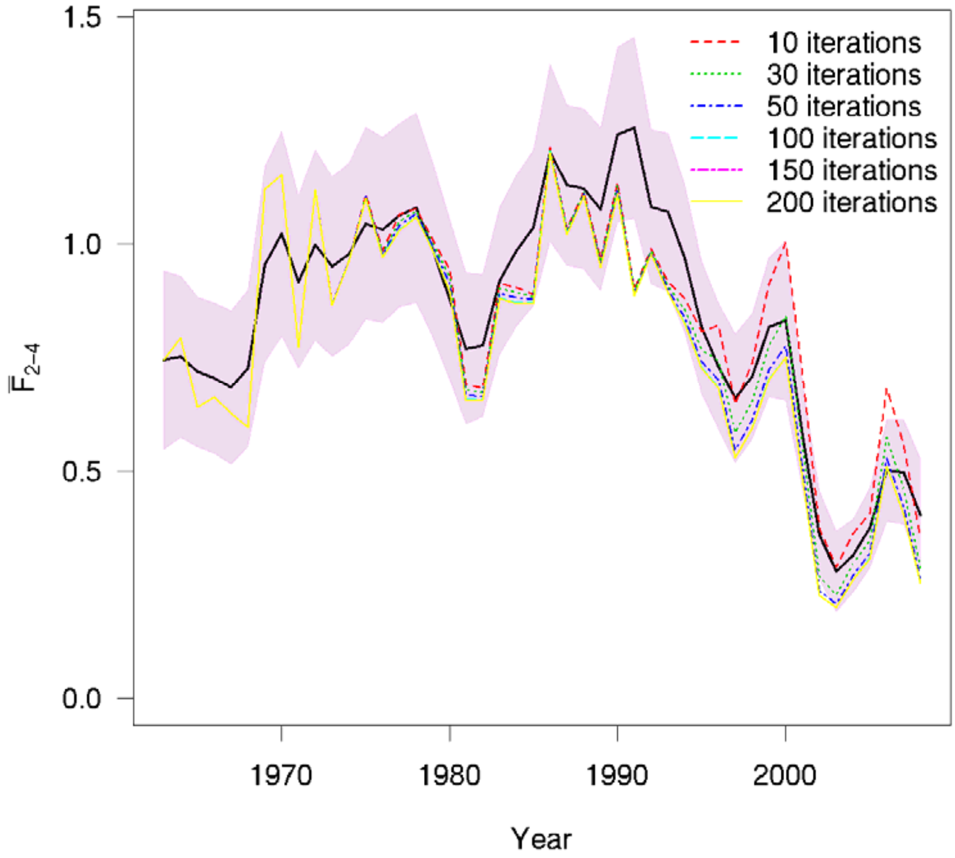

Figure 5.5.2. Average fishing mortality estimated via the standard state-space model (thick black line) and corresponding 95% confidence interval (shaded areas), and by XSA with different numbers of iterations (dashed lines).

The different number of iterations in the XSA assessment results in different stock sizes and fishing mortalities, primarily for the last 10–15 years. The final year estimates of spawning-stock biomass and average fishing mortality are especially important to future predictions, and hence for management decisions.

The state-space assessment model does not iterate backwards from initially guessed survivors, so the backwards-in-time convergence is not an issue for that model. The convergence of the state-space model is a standard gradient criteria used in maximum likelihood estimation. The convergence in the final year is no different from the convergence in the first year.

Figure 5.5.1 shows that the unconverged XSA run with 10 iterations is the most similar to the SAM estimates for SSB. That is, *if* the SAM run is the truth, *then* continuing the XSA run to convergence is pushing the XSA estimates away from the truth. However, for North Sea haddock we cannot know what the truth is, so firm conclusions about the XSA runs are hard to reach from this evidence. This problem is addressed further in Section 5.6.

5.6 Convergence tests with simulated data

Method

The approach used with real data in Section 5.4 was extended to consider simulated data. Ten datasets were used, the details of which are summarized in Annex 4. For each of the datasets, all combinations of the following test cases were included:

d) Plus-groups in the range 7 – 10 (4 cases).

e) *q*-plateaux at ages 3 and 6 (2 cases).

f) 10 retrospective runs (10 cases).

This produced in all 10 x 4 x 2 x 10 = 800 runs. Each run was summarized by two metrics:

4) The number of iterations N_i required for convergence. Time restrictions meant that the maximum number of iterations considered in the runs was 150, so in those cases for which $N_i = 150$, the value of N_i must be viewed as a lower bound on the true number of iterations required for convergence.

5) A categorical variable D_i describing whether the iteration process moved the final-year SSB estimate towards ($D_i = 0.0$) or away from ($D_i = 1.0$) the true value, which is known as these are simulated datasets. In cases where the convergence curve for SSB is very flat, or where it *crosses* the true value, D_i was set to 0.5.

Results

Figures 5.6.1 to 5.6.3 illustrate examples of each of the possibilities for the convergence criterion D_i. Over all 800 runs, the average number of iterations was $\bar{N}_i = 101.2$, while the average convergence direction was $\bar{D}_i = 0.67$. In other words, XSA took a relatively large number of iterations to converge with these simulated datasets, and convergence had a tendency to push the final-year SSB estimate *away* from the true value.

Figures 5.6.4 and 5.6.5 illustrate the dependence of N_i and D_i on the plus-group and *q*-plateau settings. A low plus-group age (7 or 8) when combined with a high *q*-plateau age (6) results in poor XSA convergence, with $N_i \geq 150$ in many cases. The *q*-plateau age has less effect on N_i when a higher plus-group age is used.

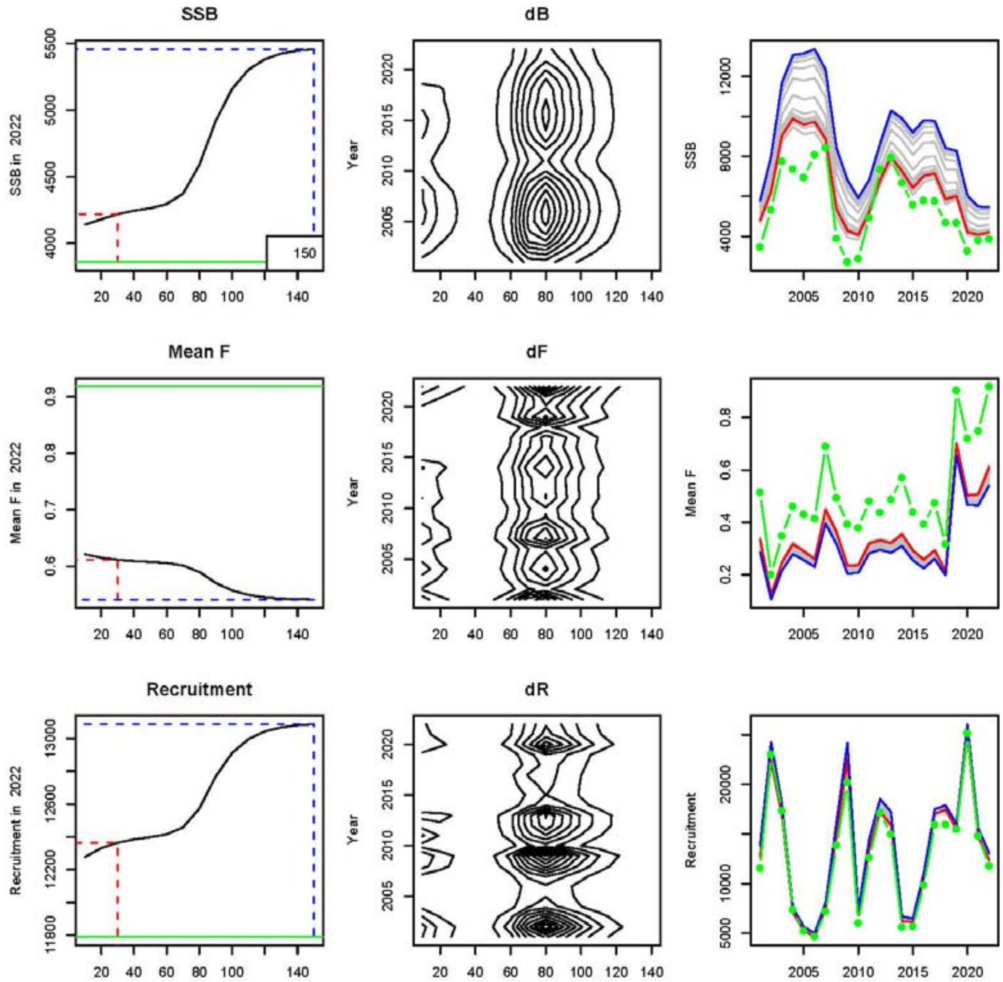

Figure 5.6.1. XSA convergence tests for simulated dataset 006 (9th retro), with plus-group at age 9 and *q*-plateau at age 3. For this run D_i = 1.0. Top row: SSB. Middle row: mean F(2–5). Bottom row: recruitment-at-age 1. Left column: relationship between the estimate for 2022 and the number of iterations run (red lines indicate 30 iterations, blue lines indicate iterations required for convergence, green line indicates the true value). The key gives the number of iterations to convergence, or 150 if convergence does not occur. Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line), converged iterations (blue line) and the true values (green lines).

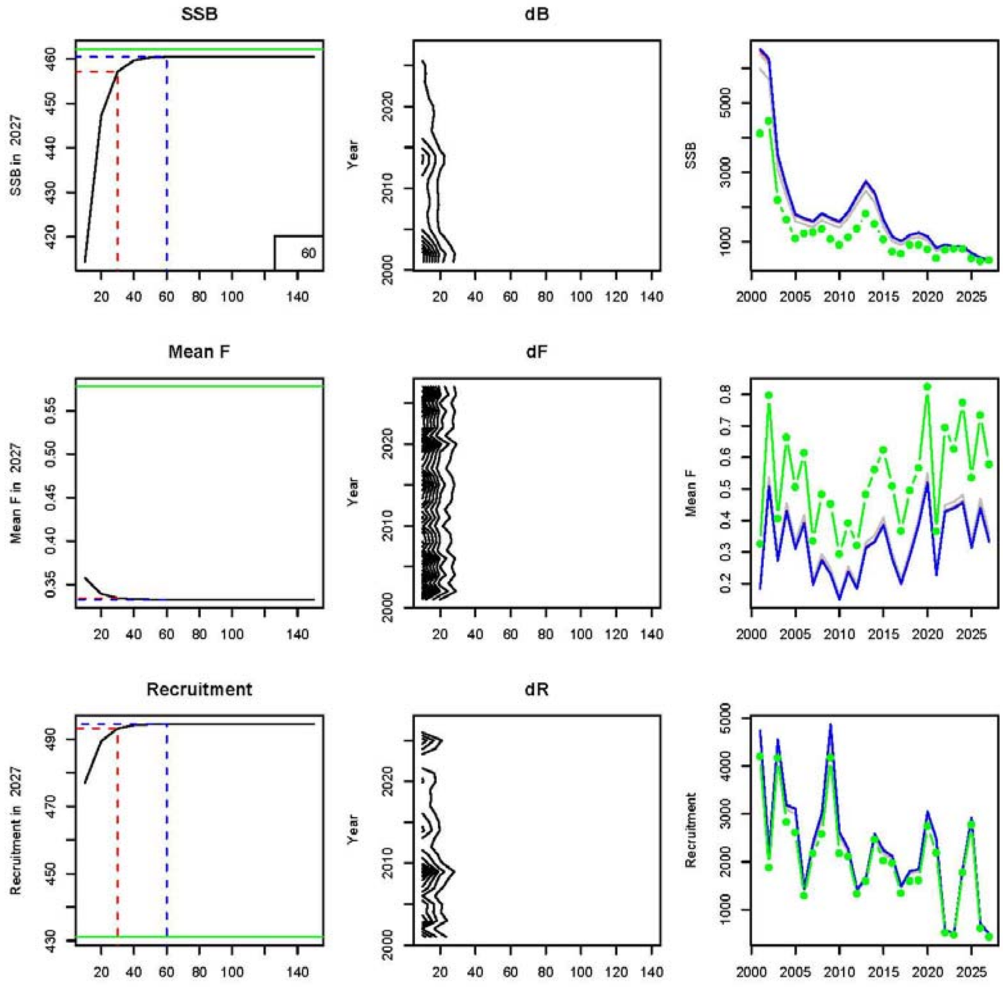

Figure 5.6.2. XSA convergence tests for simulated dataset 007 (4th retro), with plus-group at age 7 and *q*-plateau at age 3. For this run D_i = 0.0. Top row: SSB. Middle row: mean F(2–5). Bottom row: recruitment-at-age 1. Left column: relationship between the estimate for 2027 and the number of iterations run (red lines indicate 30 iterations, blue lines indicate iterations required for convergence, green line indicates the true value). The key gives the number of iterations to convergence, or 150 if convergence does not occur. Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line), converged iterations (blue line) and the true values (green lines).

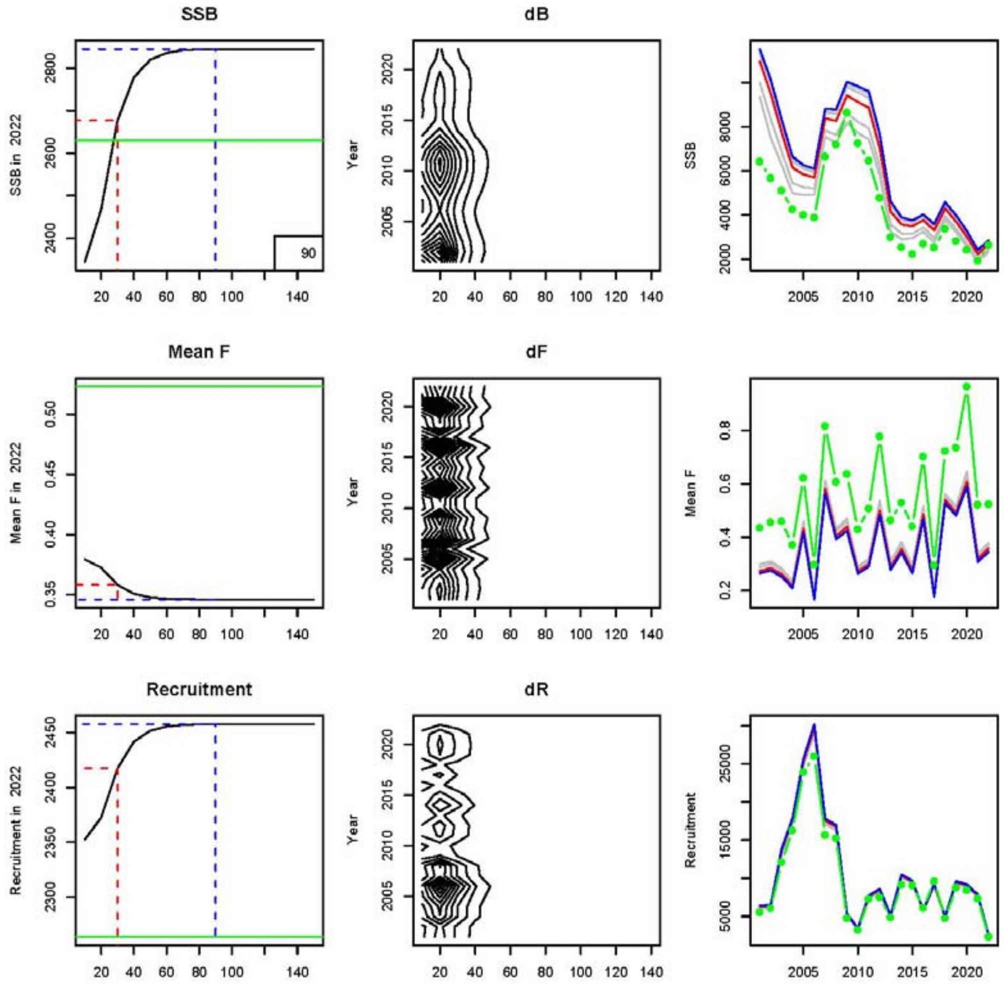

Figure 5.6.3. XSA convergence tests for simulated dataset 009 (9th retro), with plus-group at age 8 and *q*-plateau at age 3. For this run D_i = 0.5. Top row: SSB. Middle row: mean F(2–5). Bottom row: recruitment-at-age 1. Left column: relationship between the estimate for 2022 and the number of iterations run (red lines indicate 30 iterations, blue lines indicate iterations required for convergence, green line indicates the true value). The key gives the number of iterations to convergence, or 150 if convergence does not occur. Middle column: contour plot of difference between estimates over the whole time-series between one iteration and the next. Right plot: estimated time-series from all iterations (grey lines), with 30 iterations (red line), converged iterations (blue line) and the true values (green lines).

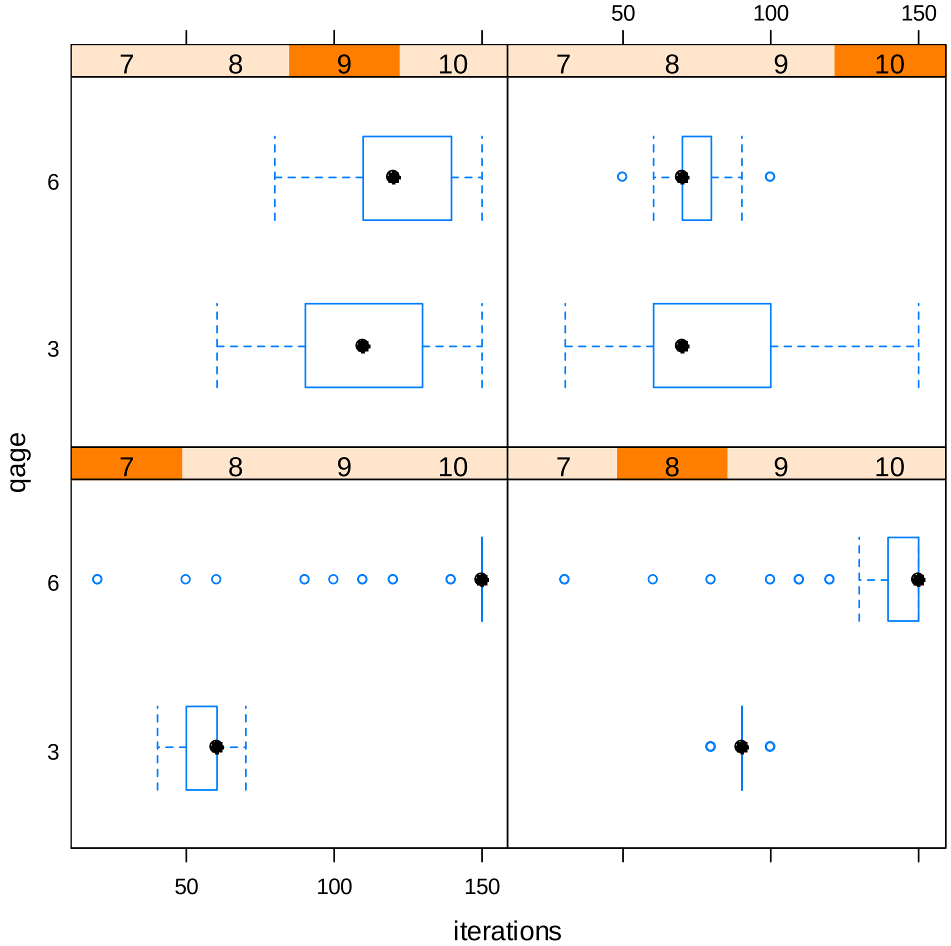

Figure 5.6.4. Box-and-whisker summaries of the number of XSA iterations required for convergence across 10 simulated datasets, each with 10 retrospective runs. The *y*-axis gives the *q*-plateau age, while the strip header for each subplot gives the plus-group age.

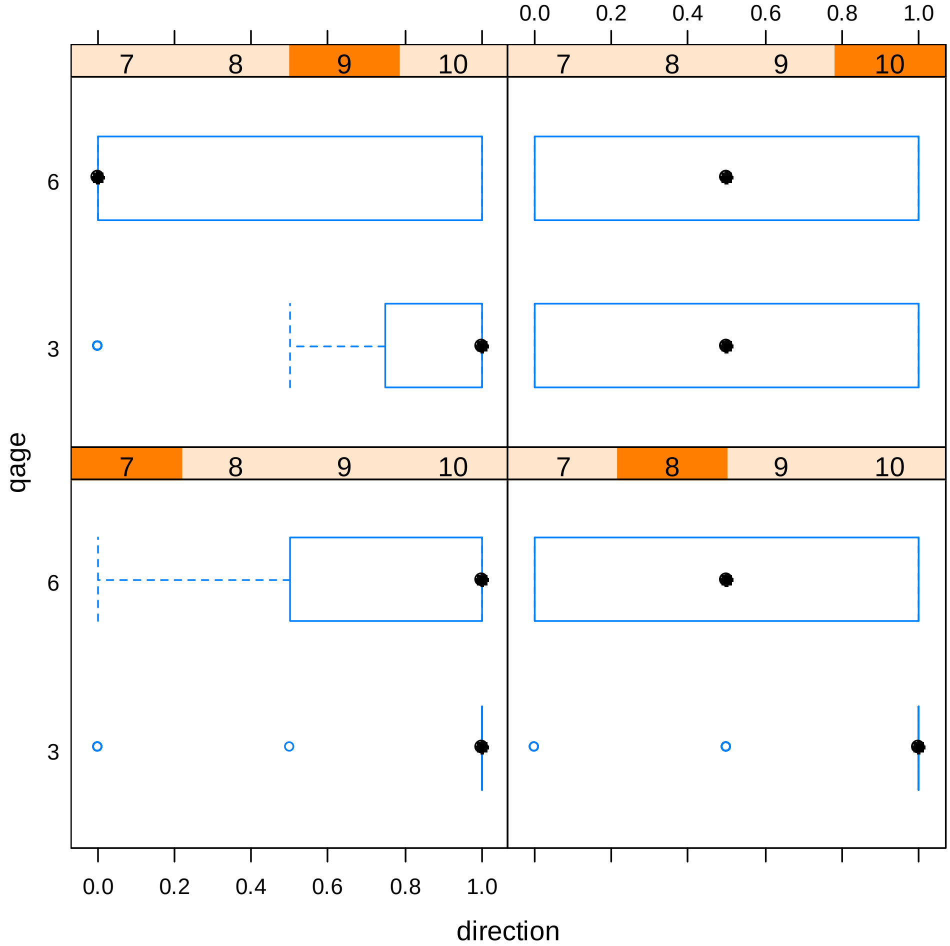

Figure 5.6.4. Box-and-whisker summaries of the convergence direction (0.0 = towards the true value, 0.5 = crossing the true value or flat, 1.0 = away from the true value) across 10 simulated datasets, each with 10 retrospective runs. The *y*-axis gives the *q*-plateau age, while the strip header for each subplot gives the plus-group age.

For these simulated datasets, a *q*-plateau age of 3 leads to a strong tendency for convergence to move away from the true value for plus-group ages 7 to 9: this effect is not seen for plus-group age 10.

A low *q*-plateau age means that the number of parameters that XSA needs to estimate is reduced, as one *q* estimate is then used for a number of older ages. The effect of this reduction with these datasets is dependent to a degree on the plus-group age chosen, but the overall tendency is for parameter reduction via a low *q*-plateau age to a) reduce the number of iterations required for convergence, and b) cause that convergence to push SSB estimates further away from the true value.

5.7 Conclusions

The convergence properties of XSA are difficult to determine. It is generally true with model fitting that there is a trade-off between the number of parameters and the efficiency of the model fit. This appears to be the case for XSA when applied to some simulated datasets: reducing the number of catchability parameters to be estimated also reduces the number of iterations to convergence, but at the cost (in general) of poorer estimates. The increase in the number of iterations certainly occurred when the number of parameters to be estimated was increased for North Sea whiting (see Section 5.4). On the other hand, the comparison of XSA and SAM assessments for North Sea haddock could indicate that the XSA assessment for that stock has *both* slow convergence *and* poorer estimates as more iterations are run – which is not a trade-off at all. This perception does depend, however, on treating the SAM assessment as the truth, which cannot really be justified.

Firm conclusions on these issues are difficult. This chapter has at least shown that assessments and subsequent advice can be very sensitive to essentially *ad hoc* choices about iterations (much as Section 4 demonstrated the importance of *ad hoc* choices about shrinkage). In theory any model should be run until it has converged, but the simulation analyses suggest that (in many cases) this convergence may indeed produce assessments that are further from the truth. Whether this happens or not depends on a number of data and modelling issues, and in reality we cannot test for this effect as we don't know what the truth is.

We conclude by returning to the questions posed at the start of this section:

1) Should XSA runs be iterated to numerical convergence in all cases?
   - No. We have demonstrated that XSA convergence may push assessments further from the truth. However, we cannot suggest suitable iteration cut-off points either, because of the possibility that convergence is improving estimates (as it does for around 33% of simulated cases). It is impossible to tell which of these possibilities is occurring for any given XSA assessment without knowing the true population structure.

2) Are assessment Working Groups applying consistent convergence criteria?
   - No. Convergence criteria appear to have evolved over time in different ways for different Working Groups, and there is no consistency. Point 1 above suggests that consistency is not possible in any case.

3) What XSA run settings are likely to affect convergence?
   - We have considered *q*-plateau age and plus group age, and both of these appear to affect convergence in both real and simulated datasets. This should not be considered an exhaustive list, however.

WGMG considers that it is essential to determine the convergence characteristics of any assessment. WGMG finds further that, in cases where the convergence of the method used is considered problematic, it is useful to explore alternative methods. In particular, if the XSA model has very slow convergence properties *and* convergence leads to very different stock perceptions, then serious consideration needs to be given to the possibility that the XSA model may not be suitable for that stock. The state-space assessment model is a valid alternative based on a simple maximum likelihood approach, but other potentially equally valid models exist.

Recommendations:

1) FLXSA output should include the number of iterations taken to reach the solution.

2) Convergence behaviour of XSA (and indeed all models) should *always* be checked.

3) Slow convergence and/or high sensitivity to the number of iterations could indicate that XSA is not suitable for the stock concerned. Alternative mod-

els that do not rely on *ad hoc* assumptions and algorithms should be explored for these cases.

6 Review of environmental information in assessments and advice (ToR a6)

Several authors have explored the potential benefits and pitfalls of incorporating environmental information to improve fisheries management. For example, Cochrane and Starfield (1992) found that there was the potential for increasing average catches of the highly productive South African anchovy stock by up to 48% if very precise short-term predictors of recruitment could be found. This is because substantial fishing on incoming recruits (age 0) can occur before the results from the first acoustic survey on these fish becomes available, so that in the absence of additional information prior to the survey, TACs are necessarily more conservative to account for the possibility of poor recruitment. De Oliveira and Butterworth (2005) conducted a simulation study on the same stock using Management Strategy Evaluation (MSE; e.g. Kell *et al*., 2007, ICES-SGMAS 2008) to investigate how potential benefits are related to the proportion of variation in recruitment explained by the environmental index. They also investigated the extent to which these benefits are compromised by uncertainties related to the degrees of freedom effect (over-fitting data), the selection of explanatory variables (danger of spurious correlations) and errors in the values of explanatory variables (including measurement error). They used the environmental indices as recruitment predictors that adjusted the TAC depending on whether these indices indicated recruitment to be in the top or bottom third of the distribution of possible recruitment values. They found that environmental indices need to explain at least 50% of the total variation in recruitment (coefficient of determination, $r^2>0.5$) before management strategies showed any benefits in terms of risk and average catch. For lower r^2, performance was *worse* in terms of average catch when incorporating the environmental index than when ignoring it.

Basson (1999) used a Monte Carlo simulation approach to investigate whether there were likely to be any gains from incorporating an environmental factor into management. She found that uncertainty could only be reduced if the environmental factor could be well predicted, and if the interaction between the environmental factor and recruitment was strong. Furthermore, the magnitude of any gains depended on the life history and fishery parameters associated with the stock: for low productivity resources (e.g. gadoid-like species), there were no gains (in terms of either average yield or conservation) when the environmental index was incorporated in the short-term prediction of recruitment, but gains were possible due to changes in fishing mortality reference points when the environmental index could be well predicted. However, there were situations where one could do worse by explicitly incorporating the environmental factor because of poor prediction capabilities, such as, for example, when predictions based on temperature are out of phase with the actual recruitment series.

Walters (1989) used stochastic dynamic programming to investigate expected improvements in management performance resulting from the use of recruitment forecasts, and found that improvements depended strongly on the average productivity of the stock concerned, and the flexibility of the in-season regulatory system used to manage that stock. For example, productive stocks managed with inflexible annual quotas showed large improvements (30–50%) in average yield if perfect preseason forecasting was practical, but unproductive stocks showed only modest improvements regardless of the in-season regulatory system used. Walters' findings are consistent with those mentioned above.

There are relatively few examples worldwide of environmental indices actually being used to manage fish stocks (Barange 2001, 2003; Barange *et al.*, 2009). The classic example is that of California sardine (Deriso *et al.*, 1996, PFMC 1998), where the HCR used for management includes a target level of fishing mortality that is a function of an environmental variable (the average sea surface temperature (SST) at Scripps Peer, La Jolla, for the three seasons preceding the year for which the catch limit is needed). The justification for the use of the environmental variable is that MSY and B_{MSY} depend on habitat area, and therefore monitoring habitat area through an appropriate proxy (i.e. sea surface temperature (SST), which has been correlated with sardine productivity, Jacobson and MacCall, 1995) helps anticipate periods of high and low productivity, so that management can be adjusted appropriately (Jacobson *et al.*, 2005). The environment-recruit relationship for California sardine, based on sea surface temperature (SST), was one of the few such relationships that Myers (1998) noted was confirmed when new data became available.

A contrasting example where an environmental index was not used successfully in fisheries management is provided by Bay of Biscay anchovy. Borja *et al*. (1998) found that an upwelling index was significantly correlated with annual recruitment of Bay of Biscay anchovy for the period 1967–1996, explaining some 59% of the variability of recruitment. The corresponding relationship was subsequently used as a basis for predicting recruitment of age 1 fish in 2000, which led to the SSB estimate based on this prediction falling below the precautionary SSB level, and the TAC for 2000 being halved as a result (ICES 2000, 2001). However, subsequent information indicated that recruitment had been substantially underestimated, leading ICES to conclude that the upwelling index had only limited use as a predictor of absolute recruitment (ICES-WGHMSA 2001, 2002). The practice of using the upwelling index to modify TAC levels for Bay of Biscay anchovy was subsequently abandoned.

Factors that may have contributed to the successful use of an environment-recruit relationship in fishery management for the California sardine, and the failure of that for Bay of Biscay anchovy include the following:

a) The environment-recruit relationship for California sardine was verified when new data became available (Myers 1998), whereas for Bay of Biscay anchovy, environment-recruit relationships that were initially thought to be strong (r^2 around 70%, Allain *et al*., 2001) were later shown to break down (r^2 around 30%, Uriarte *et al*., 2002, De Oliveira *et al*., 2005). The problem here is one of the appropriate selections of explanatory variables, the danger being that if datasets have few large and small observations (i.e. few data contrast), then it is likely that there will be an environmental index that correlates with the data (Hilborn and Walters, 1992). In this context, Myers (1998) found that the proportion of published environment-recruit correlations that were verified when tested with new data were low. A thorough selection of explanatory variables requires information independent of that used to develop the regression, or alternatively a cross-validation approach such as splitting the data in half, selecting variables on the basis of the first half, and providing the predictive relationship (if still justified by the restricted data) by regression fits to data in the second half (De Oliveira and Butterworth, 2005).

b) The HCR for California sardine that incorporated an environmental variable was selected from among several HCRs after simulation testing along the lines of Management Strategy Evaluation (Kell *et al.*, 2007, ICES-SGMAS 2008), a process that the TAC-setting procedure for Bay of Biscay anchovy did not undergo at the time (although subsequent simulation testing has been undertaken, De Oliveira *et al.*, 2005). In this context Basson (1999) notes: "Simulation studies are cheap. In the context of fisheries management as opposed to pure scientific research, it is therefore crucial that advocates of the incorporation of environmental factors into fishery management and prediction procedures check whether any gains are likely to be made from expending vast amounts of long-term effort (and funds) to gain the kind of understanding required for their proper incorporation." Testing the utility of indicators in management simulations, including the development of implementation frameworks that are informative and robust to errors, is considered an important requirement prior to the formal application of such indicators (ICES-WKEFA 2007).

Concern over the best way to use environmental indicators and drivers in the provision of fisheries management is by no means new. SGGROMAT was convened for two meetings (ICES-SGGROMAT 2002, 2004) to address this very question. It was an unusual group for the time, in that it was attended by both stock assessment scientists and ecosystem process modellers, and it led to much fruitful discussion.

A workshop on the integration of environmental information into fisheries management strategies and advice (WKEFA) was held in 2007 (ICES 2007). The workshop considered a number of case studies (some of which have already been mentioned above) involving a wide range of demersal and pelagic stocks, as well as some generic stock simulations. The case studies were used to discuss and formulate generic concepts to improve fisheries management strategies and advice. The workshop considered that incorporating observed short-term changes (e.g. growth, maturation) can improve management, but only if error in the information is included in an appropriate manner. The discovery and subsequent breakdown of environment-recruit relationships highlights the need to test the utility of such relationships, as discussed earlier. Incorporating medium-term changes requires a different approach. Where explicit relationships exist, the mean of stochastic projections can be modified appropriately, but where they do not appear to exist or there is no basis for predicting environmental drivers into the future, advice should be based on scenario testing along the lines of Management Strategy Evaluation (ICES 2007, 2008).

A general recommendation of WKEFA (ICES 2007) was that, in the light of climate change, rather than assuming that a mean derived from the recent past best represents future values for any given parameter, trends should be considered and attempts made to estimate these. This calls for the development of tools that evaluate the estimates of current values and trends in the presence of both measurement and process error. A number of specific recommendations were also derived (ICES 2007), relating to robustness to regime shifts, the influence of changes in habitat on measurements and carrying capacity, the influence of changes in growth, maturation and recruitment (due to the environment) on short- and medium-term advice, and the use of multispecies models primarily for simulation testing.

There are many examples of analyses that attempt to identify environment-recruit relationships (e.g. Mackenzie and Köster, 2004, Megrey *et al.*, 2005), but these studies have tended to consider such relationships outside the stock assessment model (i.e.

the stock assessment model provides estimates of recruitment, which are then used in a separate modelling exercise to identify environment-recruit relationships). Maunder and Watters (2003) found that an integrated approach, whereby the environmental variable is directly incorporated into the stock assessment model by assuming recruitment is proportional to the environmental variable and allowing for additional temporal variation, led to superior model performance compared to correlating model estimates with the environmental variable outside the estimation procedure.

Cases where environmental variables are directly incorporated into stock assessment remain few, but are increasing (Punt, 2008). Examples are given by Maunder and Watters (2003) for the snapper stock in Hauraki Gulf-Bay of Plenty, New Zealand, and by Schirripa (2007) for the assessment of the sablefish resource off the continental US Pacific coast. Mackenzie *et al*. (2008) provide an example for Baltic Sea sprat whereby the environmental variable (North Atlantic Oscillation), although not directly incorporated into the stock assessment itself, is used in the short-term forecast to predict key advisory-related variables such as SSB and landings. This is because the environmental variable allows a prediction of recruitment to be made earlier than the annual assessment meetings, and can be used to replace the usual short-term forecast assumption of geometric-mean recruitment.

7 Influence of uncertainty in age–length keys (ToR a7)

7.1 Origin and issues with age data

The conversion of length distributions into age compositions is known to be an issue for the stock assessment of roundnose grenadier because of 1) the lack of sampling, and 2) the difficulty of reading otoliths for this fish. The current age–length key (ALK) is the result of 2713 readings of otoliths over the periods 1996–1997 and 2002–2004. In France, samples were taken from commercial landings according to the EU Data Collection Regulation. All data comes from ICES Divisions Vb, VI and VII. Spain also collects the same information but only in Hatton Bank (Subarea XII and Division VIb). Faroese Islands and Scotland do not have any age data for this stock.

An ICES workshop (WKARRG) was convened in 2007 with age readers from different countries and with different levels of experience in age reading (ICES-WKARRG 2007). This workshop aimed to review the age reading techniques, to agree age determination criteria, and to evaluate precision in age reading through comparisons of results of age reading from the same set of otoliths.

Roundnose grenadier is aged by counting the rings in otoliths and scales. Age estimation with scales is higher for younger fish but much lower than with otoliths for older fish. Whole otoliths can be read only for very small individuals. Some marginal increment analyses have also been done and have suggested the rings in the otoliths are formed annually. Another approach may come in future from otolith weights which are highly correlated to age. Weight seems to be a good predictor of age but has not been used in assessment so far. Otolith surface may also be another good predictor. A list of relevant papers on this topic is available in the WKARRG report (ICES-WKARRG 2007).

Currently, no direct validation of the age estimation has been done. This process is used to estimate the accuracy of an age estimation method in order to demonstrate the age reading is sound and based on fact (Panfili *et al.*, 2002). This means the ageing of this species from visual counting of rings may add some bias if for example, some rings would be systematically unseen and remaining unaccounted.

During WKARRG, it was shown that the agreement between readers for the same set of otoliths was low which suggested age readers may require specific training to reduce the risks of errors on age estimation. Proposals were made to establish the basis of a manual for age reading of roundnose grenadier. It also appeared that similar work should be done regarding potential bias when measuring preanal fin length. The tail of this fish is very fragile: therefore pre-anal fin length is measured instead of total fish length.

The WKARRG workshop did not explored the implication for assessment of the low agreement on age reading but mentioned that *"considering roundnose grenadier can live up to 70 years, the precision and the bias of age estimates should be evaluated according to the needs for the assessment."*

The ALK used at the ICES Working Group on the Biology and Assessment of Deep-Sea Fisheries Resources (WGDEEP) is aggregated and applied to each year of the time-series (1990–2008) used for the assessment of this stock (Table 7.1.1). This approach has been taken mainly because of the lack of otolith readings despite the fact that a substantial decrease (5.5 cm) of the average size of individuals has occurred during that period. The average weight was around 2030g in 1990 and was around 850g in 2008 (Pawlowski and Lorance, 2009). This may be an effect of the harvesting of this stock but may also be the result of the combination of the evolution of fishing depths since the beginning of the fishery (because the depth distribution of these fish are related to their age.)

Table 7.1.1. Age length key used for the Roundnose Grenadier assessment in Vb, VI, VII.

Preanal length (cm)	Age																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1		2																										
2		18	3																									
3		1	4	1																								
4			3	11	4	1																						
5				3	4	18	1																					
6			2	3	4	7	5	4	2																			
7			1	1	7	5	4	7	4	3		1	2															
8				2	3		2	3	4	3	4	3	1	2			1											
9								2	1	6	12	5	6	1	3		2											
10									1	4	5	6	7	8	6	3	1	2			5			1				
11										3	2	1	3	4	5	6	2	5	6	2	5	3	2					
12									1		4	5	1	4	5	7	9	7	8	5	3	6		2			1	
13													4	3	4	12	13	15	13	18	5	9	6	1	5	7	3	1
14												1		1	6	1	14	22	21	15	18	24	19	13	9	18	8	9
15													2		3	1	13	2	28	26	34	32	37	35	32	14	2	7
16														1	2	2	5	7	22	21	35	32	46	39	34	36	27	28
17																1	5	9	11	13	29	24	32	38	39	34	38	28
18																	1	1	5	8	16	8	22	24	26	3	14	24
19																	1	2	6	7	1	8	9	19	15	8	11	19
20																	1	2		1	3	2	4	9	15	14	13	8
21																				1	3		2	4	5	6	4	
22																				2	1		1	1	2	3	1	
23																						2					1	
24																												1
25																												
26																												
27																												

Preanal length (cm)	Age																										
	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
1																											
2																											
3																											
4																											
5																											
6																											
7																											
8																											
9																											
10																											
11		1																									
12		1																									
13				2																							
14	6	2	5	6							1																
15	15	1	4	7	4	5	1	2		2					1		1										
16	23	11	18	12	4	7	3	1	1	3	2	2	1														
17	33	24	23	16	8	3	6	2	2	1		2	3	2					1	2							
18	22	18	15	12	1	8	9	2	3	4	1			2			1	3									
19	9	15	19	2	12	6	3	5	7	1	3	2	6	1	2	3											
20	6	4	3	8	14	7	4	8	6	3	2	1	3			1	2	3		1	1						
21	5	3	6	9	6	4	7	3	4		2	3	1	2	2	1	3						1				
22	4	1	5		3	4	1	2	1	1	2		1		3	1	2				1	2				1	
23		1			2		1		2	3				1	3												
24	2		1				2																				
25	1				2																						
26																											
27																											

7.2 Quantification of uncertainties in the assessment

A bootstrap analysis has been carried out to integrate the uncertainties from the age length key into the assessment. Roundnose grenadier stock assessment is generally done using SVPA within ICES, using the Lowestoft VPA95 suite under MS-DOS. This suite does not allow scripting therefore for this exercise; the separable VPA was performed with the sepVPA routine from the FLR package (FLAssess version 1.99–6) under R 2.7.2.

While the results from sepVPA were close to those obtained from the VPA95 suite, they were different especially for mortality-at-age. It has been impossible to find a way to replicate the same results. The documentation of FLAssess, the R library containing the sepVPA routine, does not contain many details on how sepVPA works or about the inputs and outputs of this routine. Residuals from the assessment were also not available. However, there no solid reasons to believe this may have changed the results and conclusions from this analysis.

An initial set of 2000 bootstrap replicates was made by resampling with replacement the ALK. The exploration of replicates showed standard deviations and means for stock biomass, numbers and fishing mortalities become stable after 500 hundred replicates. Subsequent runs used 800 replicates to reduce processing time. Each bootstrap replicate has been used to convert length distributions in age distributions and to get catch- and weight-at-age matrices and an assessment was performed for each replicate.

All assessments were carried out using landings and length distributions for division Vb and subareas VI, VII for the full period 1990–2008. All assessments used the following parameters:

- Model was run on age-groups 16 to 40, the 40 y.o. group being a plus group.
- The reference age-group was 25 y.o.
- Terminal fishing mortality F was set to 0.1.
- Selectivity factor S was set to 0.8.

The resulting catch-at-age matrix (Figure 7.2.1) shows some strong differences of average CV according to age. The lowest errors are observed for the 16–37 y.o. classes (around 12%) while the highest values are observed for the youngest and oldest ages. This is related to the lack of samples available for those classes. The CVs decrease from classes 10 to 24 (respectively 99% to 8%), are then more or less stable at 12% up to 32 y.o. then start to increase at a faster pace up to 104%. With the assessment based on the 16–40 y.o. classes, this results suggest the uncertainties generated by a small number and samples are kept minimal.

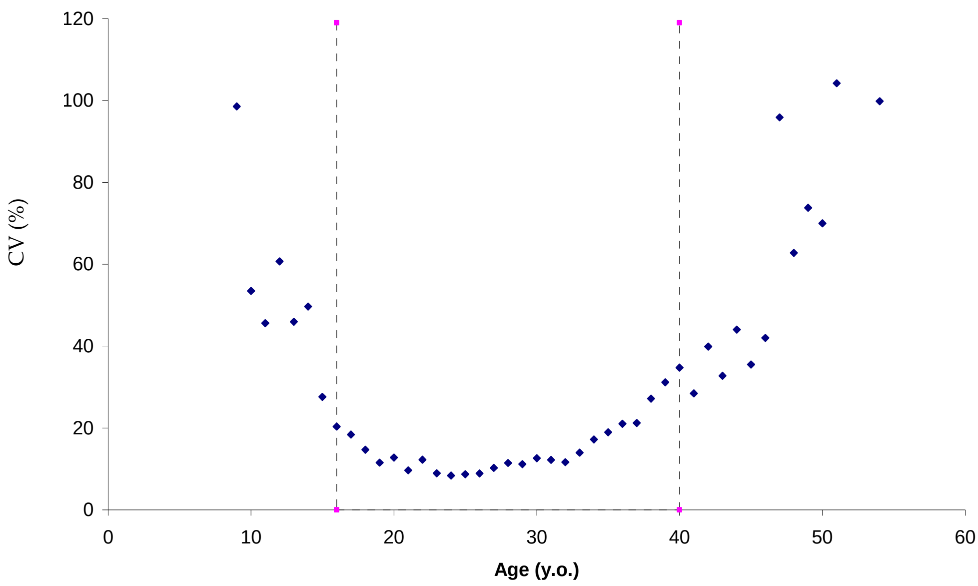

Figure 7.2.1. Average CV on catch-at-age (%) per age group.

Both biomass and numbers of individuals (Figures 7.2.2 and 7.2.3) have decreased since the early days of the fishery. In recent years, numbers and biomass seem to increase slowly. The CVs associated with both model outputs are high at 31% with however, no substantial change between years.

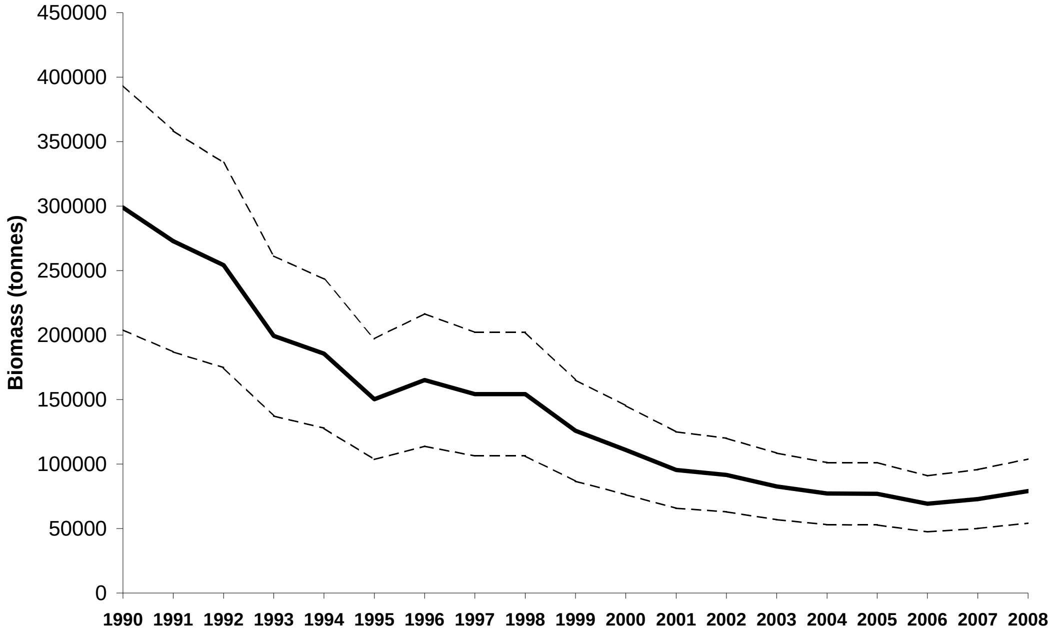

Figure 7.2.2. Evolution of biomass. Dotted lines are mean biomass +/- standard deviation.

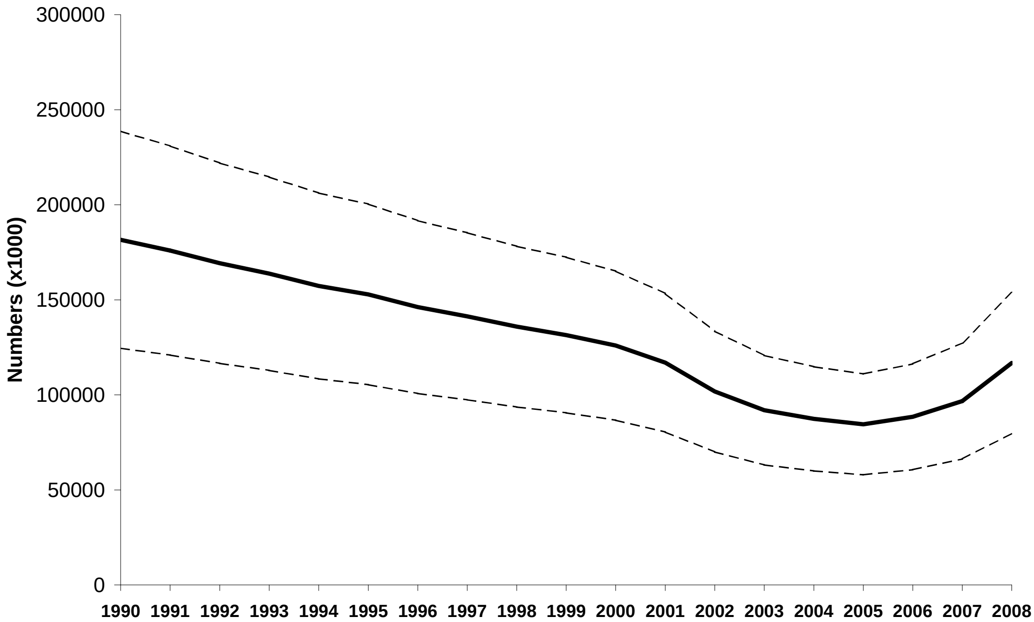

Figure 7.2.3. Evolution of number of individuals. Dotted lines are mean biomass +/- standard deviation.

Fishing mortality (Figure 7.2.4) for this stock increases from 1990 to a peak in 2001 which corresponds to the peak of landings for this fishery. F then converges towards terminal F. The CV (Figure 7.2.5) is maximal at 41% in 1991 then gradually decreases towards 0% in 2008 as this methods sets force the assessment to reach a set terminal F in the final year.

Figure 7.2.4. Evolution of fishing mortality. Dotted lines are mean biomass +/- standard deviation.

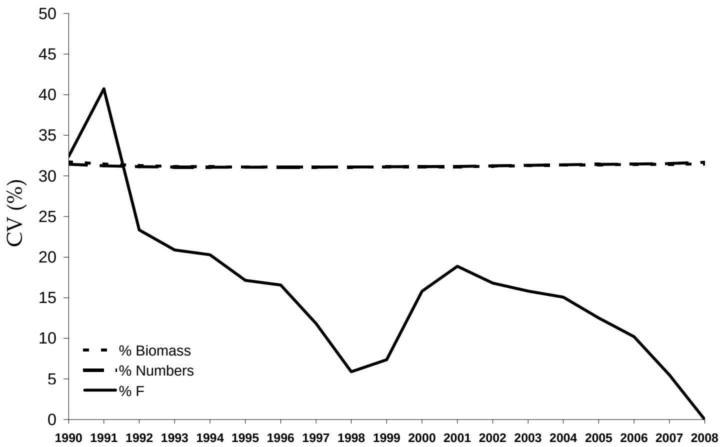

Figure 7.2.5. Compared evolution of CV of biomass, numbers and fishing mortalities from year to year.

In conclusion, the uncertainties from the ALK have an influence on the uncertainty estimates for the stock assessment, but not perhaps on the resulting management advice. The level of error is high for biomass but does not alter the general idea of a strong depletion of the stock and does not affect trends.

The level of uncertainties would pose some problems if the management of the fishery was relying on absolute numbers for example to define a TAC or a management strategy. From the analysis of the catch-at-age matrix, the assessment would probably benefit from having an ALK with higher samples for younger and older individuals. The assessment is however based on the age groups having the most samples: therefore uncertainties are kept relatively minimal.

7.3 Effects of the size of the ALK on the assessment

The current ALK was made from samplings over a couple of short periods of the exploitation of this stock despite evidence of strong changes in the length distribution over time. This raises the question of whether the noise generated by using disaggregated ALK (a separate ALK year by year) with fewer samples that would maybe reflect more those changes over time than a full aggregated ALK applied blindly to a long time-series.

But fewer samples will also induce more uncertainties. Therefore, the choice of using aggregated age data should probably be influenced by setting a balance between the level of uncertainty associated with the small number of samples and how annual ALKs may reflect the change in age/length structure in the population.

Due to time constraints, it was not possible to test the effects of a disaggregated ALK but the effect of the number of samples in the ALK over the assessment was evaluated.

With the same parameters as in the previous section, a second series of runs was made by subsampling the ALK and making smaller bootstrap replicates with 2713 (actual number of samples), 2000, 1500, 1000, 500 and 100 samples. As before, 800 replicates were done for each set of subsamples. Again, those replicates were used to convert length distributions in age distributions and to get catch- and weight-at-age matrices and an assessment was performed for each replicate.

Having smaller ALKs does not change the general shape of the outputs from the assessments in terms of mean biomass, numbers and fishing mortalities. It however affects the respective CVs. The evolution of CVs through years is not different than for the case of the use of a full ALK but the amplitude of errors naturally increases the smaller the ALK is (Figures 7.3.1 to 7.3.3).

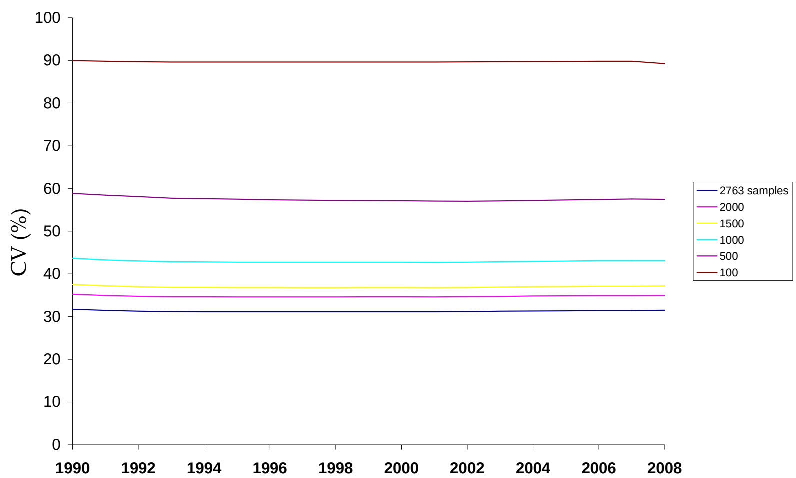

Figure 7.3.1. Evolution of CV for biomass per year and per number of samples of the ALKs.

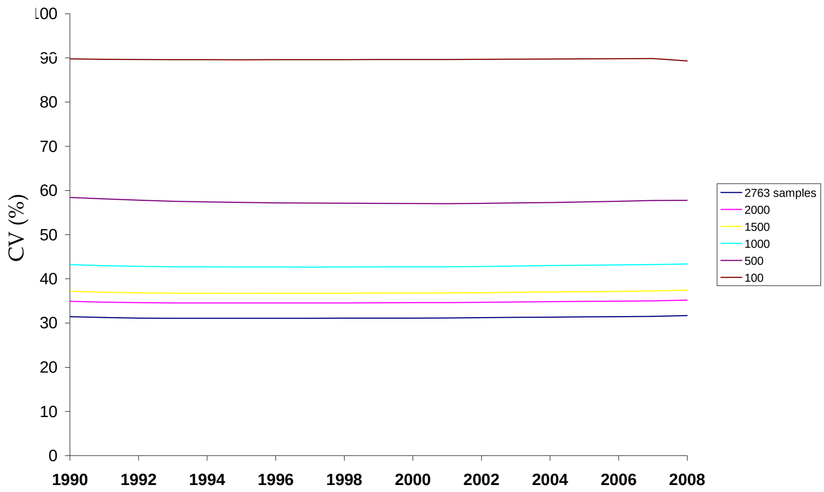

Figure 7.3.2. Evolution of CV for the numbers of individuals per year and per number of samples of the ALKs.

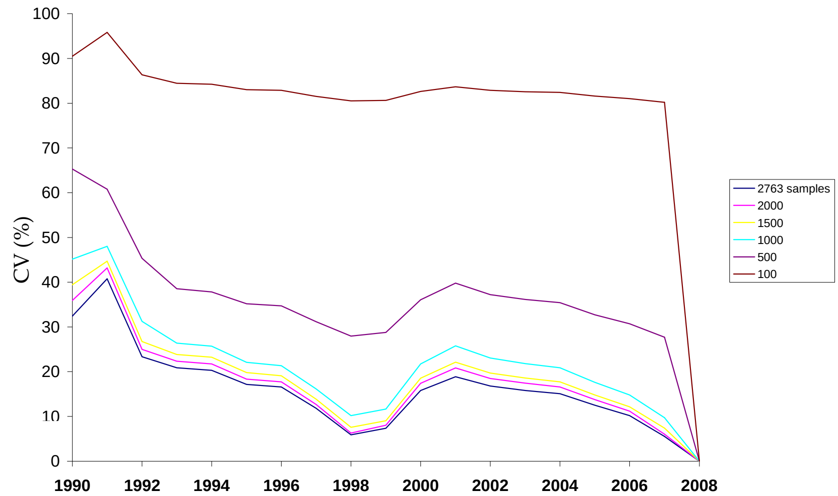

Figure 7.3.3. Evolution of CV for biomass per year and per number of samples of the ALKs.

From 2001 to 2006, according to the WKARRG report, the number of aged grenadiers per year at the IFREMER laboratory was between 480 and 1000. From the results of the current analysis, the maximum CV using yearly ALKs would be between 44 and 59% for biomass (against 32% with a full ALK) and 48 and 65% for fishing mortalities (against 41% for the full dataset) which is substantially higher in all cases.

7.4 Recommendations

- Using smaller annual ALKs for the Roundnose grenadier assessment, considering the level of annual sampling, may add more uncertainties than benefits. Therefore the assessment should probably use the entire ALK rather than disaggregating the ALK into separate subsets year by year.
- However, the effects of using annual ALK should be studied and particularly, whether using annual ALK from one year to another would reflect in some way the observed changes in the population structure since the be-

ginning of the fishery. Direct age reading could limit uncertainties but does not appear to be easily achievable considering the difficulties of age reading for this species

- With an ALK with 1500 or more samples, outputs start to have some similar levels of uncertainties than those observed for the full ALK. This value might be seen as a compromise between uncertainties and using smaller annual ALKs. For less than 1500 samples, the uncertainties generated become very high.
- The current ALK lacks individuals less than 10 and over 37 years old. Consolidating the ALK with those age classes may reduce uncertainties and would allow the use of other groups for the assessment than the current 16–40+ age groups.
- The potential biasing effect of measuring pre-anal fins should be evaluated as recommended by WKARRG.
- WGMG considers that age-based assessments are unreliable for this stock and suggests development on life-stage structured approaches. However, length-based approaches may be limited for this stock due to the lack of data on growth and uncertainties on age reading.

8 State-space assessment models

8.1 Setting zero variances in a state-space stock assessment model

Motivation

To illustrate the workings of the state-space assessment model, WGMG decided to investigate the effects of fixing certain variance parameters. The idea was to replicate the questionable assumption from the deterministic approaches that catches are known without any observation noise in the state-space framework and study the effects.

Summary of the state-space model

The state-space assessment model contains two parts. A process of underlying unobserved states α, here the log-transformed stock sizes $\log N_1,\ldots,\log N_A$ and fishing mortalities $\log F_{i_1},\ldots,\log F_{i_n}$. The second part of the state-space assessment model describes the distribution of the observations x given the underlying states α. Here x consist of the log-transformed catches and survey indices.

The transition equation describes the distribution of the next year's state from a given state in the current year. The following is assumed:

$$\alpha_y = \mathrm{T}(\alpha_{y-1}) + \eta_y$$

η is process error which is described in more detail below.

The transition function T is where the stock equation and assumptions about stock--recruitment enters the model. The equations are:

$$\log N_{1,y} = \log(R(SSB_{y-1}))$$

$$\log N_{a,y} = \log N_{a-1,y-1} - F_{a-1,y-1} - M_{a-1}\ ,\quad 2 \le a < A$$

$$\log N_{a,y} = \log N_{a-1,y-1} - F_{a-1,y-1} - M_{a-1} + \log N_{a,y-1} - F_{a,y-1} - M_a\ ,\quad a = A$$

$$\log F_{a,y} = \log F_{a,y-1}\,, \quad 1 \le a \le A$$

Here M_a is the age specific natural mortality parameter, which is most often assumed known from outside sources. $F_{a,y}$ is the fishing mortality. The function R describes the relationship between spawning-stock biomass $SSB_y = w_{1,y-1}p_{1,y-1}N_{1,y-1} + \cdots + w_{A,y-1}p_{A,y-1}N_{A,y-1}$ (here w are weights, p maturities) and recruitment. The parameters of the chosen stock–recruitment function are estimated within the model. Often it is assumed that certain F_a parameters are identical (e.g. $F_{A-1} = F_A$).

The prediction noise η is assumed to be uncorrelated Gaussian with zero mean, and three separate variance parameters: one for recruitment σ_R^2, one for survival σ_S^2, and one for the yearly development in fishing mortality σ_F^2.

The combined observation equation is:

$$x_y = \mathrm{O}(\alpha_y) + \varepsilon_y$$

The observation function O consists of the familiar catch equations for fleets and surveys, and ε_y of independent measurement noise with separate variance parameters for certain age groups, catches, and survey indices. An expanded view of the observation equation becomes:

$$\log C_{a,y} = \log\left(\frac{F_{a,y}}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right) + \varepsilon_{a,y}^{(\circ)}$$

$$\log I_{a,y}^{(s)} = \log\left(Q_a^{(s)} e^{-Z_{a,y}\frac{D^{(s)}}{365}} N_{a,y}\right) + \varepsilon_{a,y}^{(s)}$$

Here Z is the total mortality rate $Z_{a,y} = M_a + F_{a,y}$, $D^{(s)}$ is the number of days into the year where the survey s is conducted, and $Q_a^{(s)}$ are model parameters describing catchabilities. Finally $\varepsilon_{a,y}^{(\circ)} \sim N(0, \sigma_{\circ,a}^2)$ and $\varepsilon_{a,y}^{(s)} \sim N(0, \sigma_{s,a}^2)$ are all assumed independent and Gaussian.

The experiment

To mimic the strong assumptions behind the (semi-) deterministic approaches the variance parameters describing the catch observation noise $\sigma_{\circ,a}^2$ and the process variance describing the survival noise σ_S^2 were fixed. Ideally these variances should have been fixed at zero, but the estimation algorithm for the state-space model could not allow that in its current form, so instead they were fixed to $e^{-8} \approx 0.00034$.

This study was carried out using North Sea cod as the example, as a well tested implementation of the state-space model exists for this stock.

Results

The results outline the difference between a normal run of the state-space model where all model parameters are estimated and the run where the variance parameters were fixed.

The negative log likelihood for the model changed from 143.1 to 314.5 as a result of restricting the four model parameters (the three variance parameters describing the catch observation noise for the age groups $(\sigma^2_{\circ,a})_{a=1,2,3^+}$, and the variance parameter describing the survival variance σ^2_S).

Fixing the catch observation noise to zero resulted in the state-space fitting the reported catches very accurately. The biggest differences were seen at the youngest age class where the catch observations were considered most uncertain prior to eliminating the observation noise.

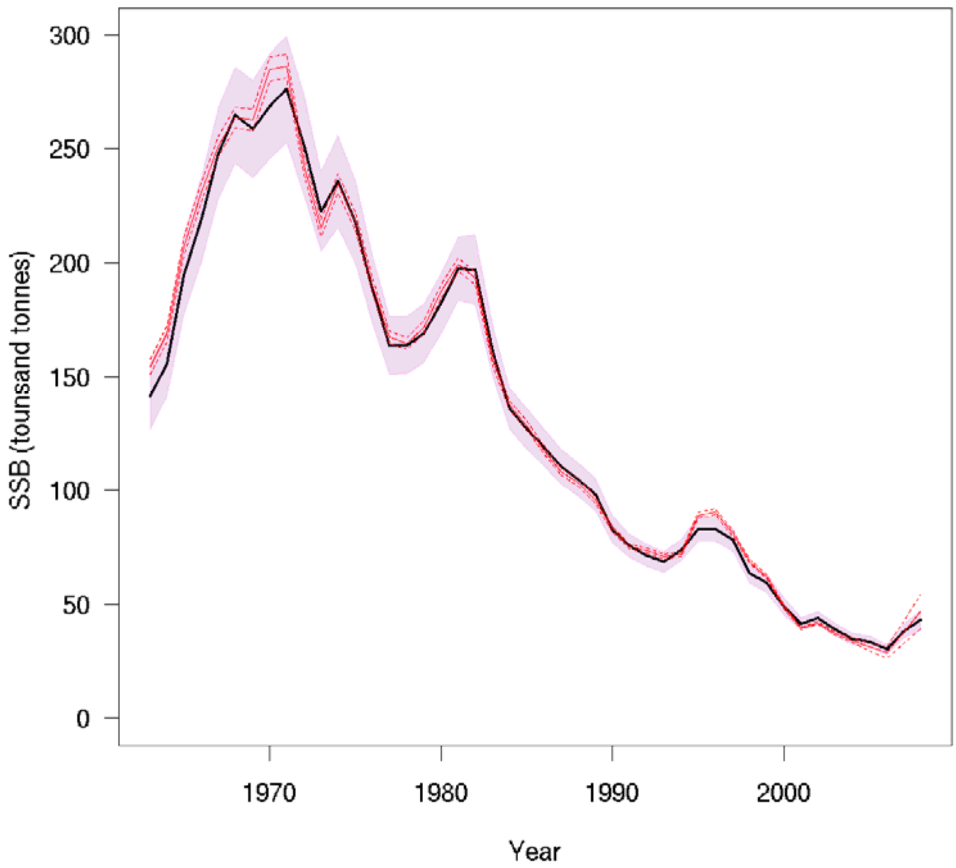

Figure 8.1.1. Spawning stock biomass estimated via the standard state-space model (thick black line) and corresponding 95% confidence interval (shaded areas), and via the variance restricted version of the state-space model (thin red line) and corresponding confidence interval (dashed thin red lines).

The time-series of the spawning-stock biomass changed only slightly (Figure 8.1.1), but the confidence intervals became unrealistically narrow (coefficients of variation less than 1%), which is an obvious consequence of pretending to have observations without noise, and consistent with the backwards convergence of the (semi-) deterministic approaches. Somewhat surprisingly there was no substantial difference in how well the survey indices were predicted. Because the age specific catchabilities $Q_a^{(s)}$ are constant in the entire data period this indicates that the relative N_a time-series are similar with and without fixing the variances.

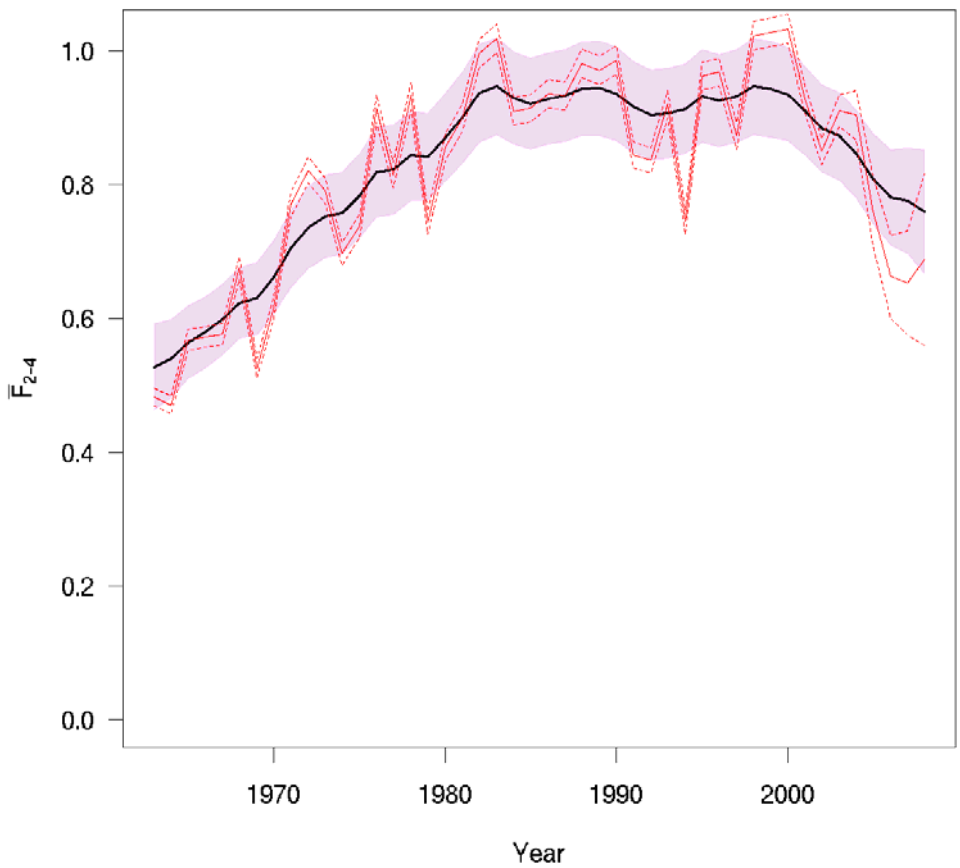

Figure 8.1.2. Average fishing mortality estimated via the standard state-space model (thick black line) and corresponding 95% confidence interval (shaded areas), and via the variance restricted version of the state-space model (thin red line) and corresponding confidence interval (dashed thin red lines).

The biggest difference was found in the estimated fishing mortalities (Figure 8.1.2). The time-series became highly variable to match exactly the assumed known catches. Furthermore, the fishing mortalities themselves became extremely well determined (coefficients of variation as low as 1% for $\bar{F}_{2-4}$).

One exception to the extremely narrow confidence intervals produced by assuming perfect catch information is the estimates in the last few assessment years. The uncertainty (standard deviation) of the last year estimates of $\bar{F}_{2-4}$ and spawning-stock biomass is almost doubled when fixing the variances compared to when estimating the variances. The state-space framework is designed to make any given year's estimate an optimal weighted average of the information from the current year, and the information from surrounding years. The assumption of noise-free catch information propagates into increased variability of the underlying process (here fishing mortality), which then makes the last years estimates more uncertain, as they almost only use information from the last year.

Conclusions

WGMG finds that the state-space assessment model is a valid alternative to (semi-) deterministic approaches. It can include observation noise in the catches and its equivalent to "year-shrinkage" is objectively estimated within the model.

If the true catch observations contain measurement noise, then the model with the same variance restrictions as the (semi-) deterministic approaches leads to fishing mortality estimates that may be too irregular, as the measurement noise is propagated here. The irregular fishing mortality estimates are consistent with the estimates seen from the (semi-) deterministic approaches.

Furthermore, assuming zero catch variance when it is not valid, causes too narrow historical confidence intervals, but a loss of precision in the final year.

9 Incorporation of survey variance in assessments

9.1 Further Extensions to the SURBA model (SURBA+)

SURBA (Needle 2008) is an age-based assessment model that can be used to estimate total mortality rates (Z's) and relative population size based only on age-based survey indices (I_{ay}, $a=1,\ldots,A$, $y=1,\ldots Y$). The basis of SURBA is a simple separable model of total mortality: $Z_{ay} = s_a^* f_y^*$, where s_a^* is the year-invariant age effect for Z and f_y^* is the year effect. Population size is modelled using the standard cohort model, $N_{a+1y+1} = N_{ay}\exp(-Z_{ay})$. Parameters are estimated using survey indices that are assumed to be related to population size via the observation equation

$$I_{ay} = q_a N_{ay} \exp(-p_y Z_{ay} + \varepsilon_{ay}),$$

where $p_y Z_{a,y}$ is the fraction of total mortality that occurs before the survey takes place, q's are parameters for the survey catchability, and ε's are observation error terms. Note that beginning of year population size, $N_{a,y}$, is projected forward to the time of the survey by applying the fraction of total mortality.

There is confounding between q's and Z's in a SURBA model (e.g. Section 4.1.2.2 in ICES-WGMG 2008). To remove this confounding, values for q's are usually supplied by the user (i.e. assumed or derived from external sources). Hence, SURBA provides population size estimates that are relative to the assumed scale of the survey q's. SURBA is a highly parameterized model, even when q values are fixed, and it is useful to control the variation in some parameter values. Shrinkage penalties have been used to reduce the between year variation in f_y's and between age variation in s_a 's. The amount of shrinkage is usually based on subjective judgment.

The above formulation of SURBA is useful for producing basic information on stock trends and total mortality rates, but it is not directly useful for evaluating management options for fisheries in the traditional sense; although it is currently used to provide trend-based management indications for North Sea whiting (ICES-WGNSSK 2009) and 3Ps cod (DFO 2009; see below) . In this Section some results are presented from preliminary investigations of extensions to SURBA that are more useful for management purposes. These involve:

1) Z model: $Z_{a,y} = M_{a,y} + F_{a,y}$, where natural mortality ($M_{a,y}$) is user-supplied and fishing mortality is modelled as a separable function, $F_{a,y} = s_a f_y$, where s_a is the fishery selectivity for different ages, which is assumed to be year-invariant. Modelling Z in terms of F and M is required to provide F-

multiplier stock projections for management advice, but does assume knowledge of natural mortality rates.

2) Control variation in f_y 's and s_a 's and reduce the number of parameters using random effects modelling approaches rather than subjective shrinkage. This provides more objective management advice.

3) Relax the separability assumption so that $F_{a,y} = s_{a,y}f_y$ and $s_{a,y}$ varies smoothly as a function of year (y). This is a more realistic assumption which may lead to more accurate management advice.

4) Use random effect approaches to model recruitment and reduce the number of parameters. This may lead to more precise management advice.

Extensions were developed for a specific stock, namely Atlantic cod (*Gadus morhua*) in NAFO Subdivision 3Ps, located off the south coast of the island of Newfoundland, Canada. SURBA has been used recently in assessments of this stock (e.g. DFO 2009).

9.1.1 SURBA+ software

A version of SURBA was described in ICES-WGMG (2008; Section 4.1.1.1) that was implemented in SAS using PROC NLMIXED. It proved to be very difficult (and perhaps impossible) to modify this code to incorporate more flexible time-varying models for selectivity (i.e. $s_{a,y}$), in conjunction with a random effects model for the F year effects (i.e. f_y 's). This software approach was abandoned in favour of an approach based on AD Model Builder (ADMB) software.

9.1.2 3Ps cod example – the survey data

The model was applied to the expanded Campelen index (including 'new' inshore strata) for 3Ps cod, for the years 1983–2009 and ages 1–12. Hence, the model provides estimates of mortality rates and trends in the size of the stock component in the survey area and at the time of the survey. This is thought to represent a large part of the 3Ps stock in total. This Campelen index is based on a stratified random bottom-trawl survey design which was expanded to include inshore strata that have been consistently sampled since 1997 (Figure 9.1.1). The expansion involved a 12% increase in the area surveyed. The survey index prior to 1997, which was based only on offshore strata, was adjusted to account for the new inshore strata. These details are described in as yet unpublished supporting DFO documents for the 2009 assessment.

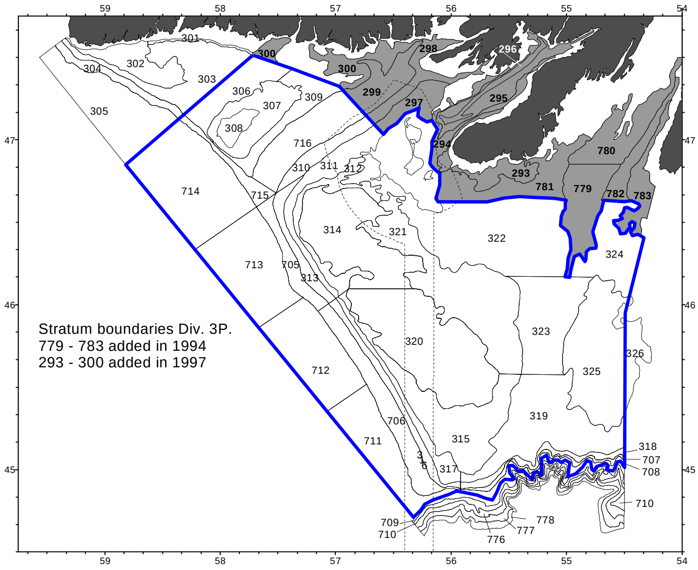

Figure 9.1.1. Survey stratification scheme in NAFO Subdivision 3Ps. "New" inshore strata are shaded.

The survey time-series of mean number per tow are illustrated in Figure 9.1.2. There is considerable interannual variability of survey catches for cod - much of which is not related to changes in cod abundance. We refer to such stock-independent variations as year effects. However, since the addition of the inshore strata in 1997, the interannual variability of survey indices have been much lower.

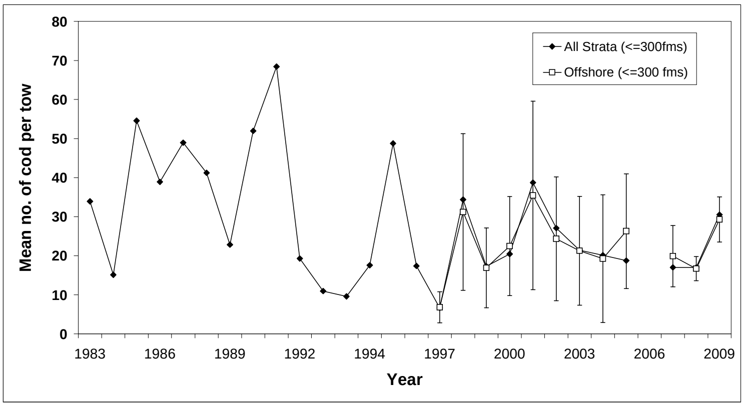

Figure 9.1.2. Strata-area weighted average mean number per tow for cod in NAFO Subdivision 3Ps.

Recent survey age compositions are shown in Figure 9.1.3. Survey catches tend to be highest at age 3, and generally decline for older ages. Note that survey catches in 2009 for many cohorts were greater than the values in 2008. This suggests that either the 2008 index was a "low" year effect; the 2009 index was a "high" year effect, or some combination of the two.

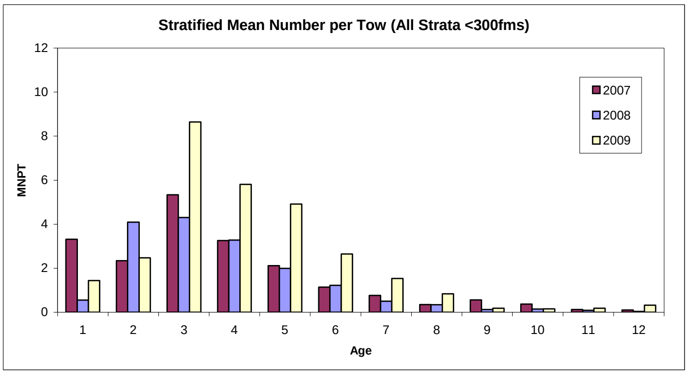

Figure 9.1.3. Survey mean number per tow (MNPT) at age for cod in NAFO Subdivision 3Ps.

Changes in the survey age composition over time are illustrated in Figure 9.1.4, using a "SPAY" plot.

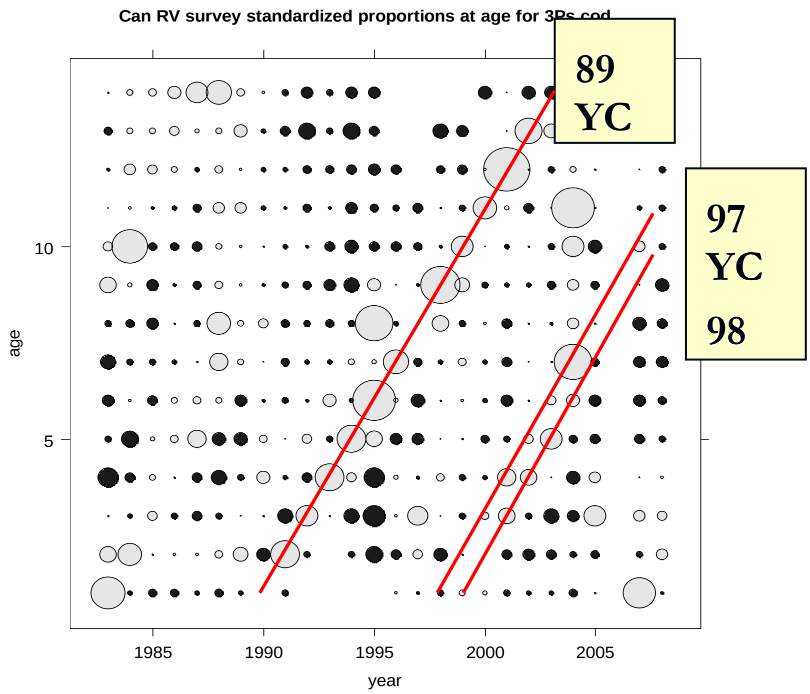

Figure 9.1.4. Standardized proportion at age per year plots ("spay"; FLEDA library within FLR) based on mean number per tow for cod in NAFO Subdivision 3Ps.

Clearly the DFO 3Ps survey data have historically been quite noisy and it is a challenge to infer trends in stock from these data. There are several possible explanations for the high variability which are available in the DFO research reports for this stock, and are not repeated here.

Cod in 3Ps have long been recognized as consisting of a complex of migratory substocks that are not all equally exploited by fisheries. It is thought that the majority of the stock is in the DFO survey area at the time of the survey (Winter-Spring); however, it is known that some stock components are not "available" to the survey, and the relative size of these components has likely varied over time. These inshore components could be heavily exploited and yet not surveyed by the DFO survey. There are additional fixed gear surveys in the inshore but it has been difficult to combine these sources of information into a single model for the stock as a whole. Commercial catches for the larger offshore component are uncertain because much of the catch occurs in the inshore in summer when that component is possibly mixed with inshore components.

For these reasons, and others, an assessment model for the entire 3Ps stock has not been accepted (or even proposed) for the last few years. In recent assessments total mortality rates and trends in offshore stock size have been inferred from survey results. The SURBA model has been used to help filter out the year-effects noise in the 3Ps index from the stock signal.

9.1.3 Model 1: Random walk for fully recruited fishing mortality

The fishing quotas for 3Ps cod have not varied much from year to year, except for 1994–1996 when there was a moratorium of commercial fishing. In addition, relatively few cod less than age 4 are landed in 3Ps fisheries; that is, they are not recruitment fisheries. This suggests that we should expect fairly smooth temporal variations in fishing mortality rates. The SURBA model used in the last two 3Ps cod stock assessment included a user supplied penalty function to control between year variations in f_y's. The penalty weight was subjectively chosen, and two options were explored that produced low and high smoothing.

An alternative approach is to treat the f_y's as a simple random walk,

$$\log(f_y) = \log(f_{y-1}) + \delta_y,$$

where δ_y are independent $N(0,\sigma_\delta^2)$ random error terms. The variance σ_δ^2 could be user specified but a more objective modelling approach is to estimate σ_δ^2 and let the data decide how much smoothing is appropriate. This is easy to do using the ADMB random effects module (ADMB-RE), which uses the marginal likelihood, in which the δ random effects are numerically "integrated out", for inference about fixed effect parameters like σ_δ^2. Typically a fixed-effect (i.e. not random) mean parameter is specified to start the random walk. For 3Ps cod the random walk starts in 1983, and we restart it in both 1994 and 1997 to account for the fishing moratorium. Hence, the random walk model is

$$\log(f_y) = \begin{Bmatrix} \mu_y, & y = 1983,\ 1994,\ 1997, \\ \log(f_{y-1}) + \delta_y, & \textit{otherwise.} \end{Bmatrix} \quad 9.3.1$$

In the first analyses of this model the estimate of σ_δ^2 was very high and the empirical Bayes predictions of the f_y's were highly variable from year to year. The problem was that the model estimation tried to account for the year effect variability of the survey indices as variability in the f_y's. As a solution we decided to model the year effect variability separately. This involved extending the observation equation,

$$I_{ay} = Q_{ay} N_{ay} \exp(-pZ_{ay} + \varepsilon_{ay}), \quad 9.3.2$$

where $\log(Q_{ay}) = \log(q_a) + \tau_y$ and τ_y are independent $N(0,\sigma_\tau^2)$ random year effects.

The other model component to specify is the selectivity, s_a. To ensure f_y and s_a effects are identified we set $s_A = 1$, as is done in SURBA 3.0 (Needle, 2008). We expect that s_a varies smoothly as a function of age. A variety of parametric models have been used for this purpose; however, sensitivity to parametric assumptions is always a concern. We decided to also use a random effects approach to produce smooth nonparametric estimates of s_a. At younger ages we expect $log(s_a)$ to increase roughly linearly with age, but at older ages we expect much less change in s_a's. To accommodate this type of variation we used the following random effects model:

$$\log(s_a) = \begin{cases} 0.5\left[\log(s_{a+1}) + \log(s_{a-1})\right] + \xi_a, & a \le 7, \\ \log(s_{a-1}) + \xi_a, & otherwise, \end{cases} \qquad 9.3.3$$

where ξ_a is a normal distribution error term. This model penalizes against first-order differences at older ages, $\{log(s_{a+1}) - log(s_a)\}^2$, and will favour constant selectivity unless there is "strong" evidence in the data for a trend. The strength of the evidence is determined via the improvement in fit to the data vs. the cost of having $Var(\xi_a) = \sigma_\xi^2 \ge 0$ in the likelihood. At younger ages the penalty is $\{log(s_{a+1}) + log(s_{a-1}) - 2\ log(s_a)\}^2$ which favours log-linear selectivities; that is, the penalty is zero when s_a is log-linear in a. If log-selectivities follow a simple random walk then $Var(\xi_{a\le 7}) = 2Var(\xi_{a>7})$, and we used the relationship to account for the extra terms in the linear penalty.

We assumed q_a = 0.1, 0.3, 0.5, and 0.8 for a = 1,…,4 and q_a = 1 for a > 4. Estimates of population size are relative to these assumptions about survey catchability.

Estimates of population size (Figure 9.1.5) suggest that total biomass in the survey area increased during 1997–2001, decreased steadily during 2005–2007, but increased from 2008 to 2009. Estimates of spawning-stock biomass (SSB) increased until 2003–04, decreased steadily to 2008, and changed little in 2009.

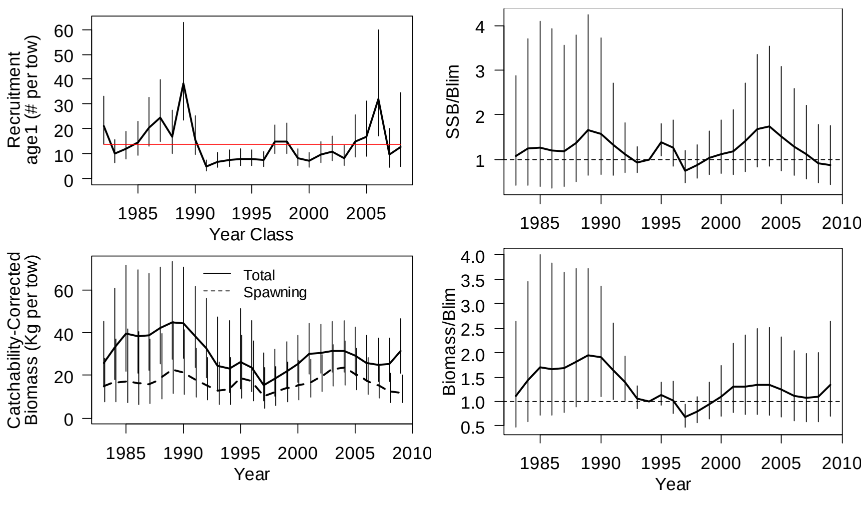

Figure 9.1.5. Estimates of recruitment (top left panel) and biomass (bottom left panel). The red line indicates the time-series mean. Estimates of SSB (top right panel) and biomass (bottom right panel) relative to 1994 values, which for SSB is the limit reference point for this stock. Vertical lines indicate 95% confidence intervals. Dashed references lines at one are shown.

The limit reference point for 3Ps cod is $B_{recovery}$ which has been identified as stock size in 1994. Estimates and confidence intervals for stock size relative to $B_{recovery}$ have been reported in recent assessments for this stock. The results (Figure 9.1.5) suggest that SSB in 2009 was slightly below $B_{recovery}$. SURBA models used in the last assessment of this stock, based on different values for q_a's, suggested that SSB in 2009 was slightly above $B_{recovery}$.

Fishing mortality rates (averaged for ages 2–4; Figure 9.1.6) increased steadily from about 0.11 to 0.43 during 1997–2007 and have declined slightly since then. Selectivity is estimated to increase log-linearly with length to a maximum of one at age eight, and is flat thereafter.

The survey catchability model (Q) includes effects to account for the inclusion of inshore strata in the 3Ps index starting in 1997 (see Section 9.1.2), and also year effects (e.g. Equation 9.3.2). The age-effects for 1997–2009 are user-supplied (see above) and the age-effects for 1983–1996 were adjusted from the 1997–2009 age effects. The adjustment factor was derived externally to the model and was based on a comparison of survey catch rates during 1997–2009 in offshore strata only compared to catch rates from offshore plus inshore strata. The year effects are predicted from the model fitting. The age effects in Q shown in Figure 9.1.7. We assumed the log year effects were iid Normal random deviates but the exponentiated predictions do not seem random. The mean and median of the year effects prior to 1994 are 1.11 and 1.12, respectively. The geometric mean prior to 1994 is 1.00. This may affect estimated trends in stock size, and this requires further investigation.

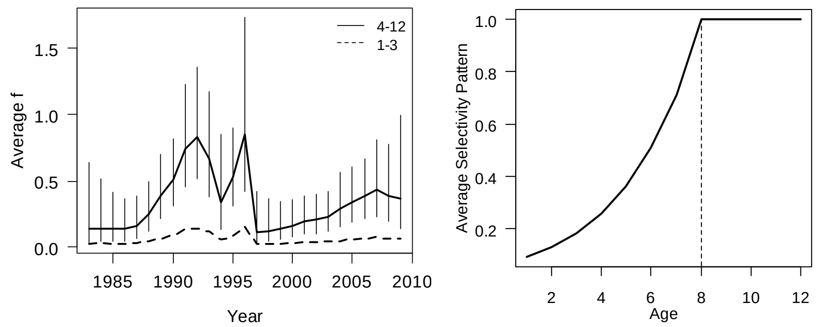

Figure 9.1.6. Left panel: Average fishing mortality for ages 4–12 (solid line) and 1–3 (dashed line). Vertical lines indicate 95% confidence intervals. Right panel: Predicted fishery selectivity. The vertical line indicates the first age that selectivity is one.

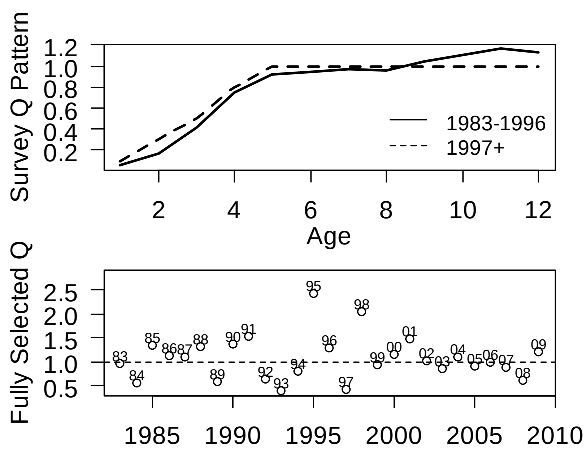

Figure 9.1.7. Age effects in survey catchability (Q, top panel) and year effects (bottom panel).

Observed and predicted age-aggregated survey indices and various plots of age disaggregated survey residuals are shown in Figure 9.1.8. There are no residuals for 2006 because the survey was not completed that year, and the index is missing. There are no residuals at ages 1 and 2 prior to 1995 because a different trawl gear was used during 1983–1995 which did not efficiently catch small cod. Data for ages 3–12 were converted based on comparative fishing studies to correspond to the current Campelen gear. There is still evidence of age×year residual patterns although these results are much improved compared to the SURBA model last used in the stock assessment for 3Ps cod. There is a slight trend in residuals at ages 1–3 which may reflect problems in the choice for survey q's for these ages.

9.1.4 Model 2: Trend in Selectivity

The separable model for fishing mortality may not be appropriate to 3Ps cod, especially before and after the fishing moratorium in 1994–1996. There were substantial changes in gear types in these two periods, with a greater fraction of landings taken with otter trawl and traps before the moratorium. Afterwards, a larger fraction of the catch was taken with gillnets which are known to have a "domed" selection pattern for cod in 3Ps.

We explored a simple mixture model in which selectivity was a weighted average of two "base" selection functions, and the weights could change from year to year. The

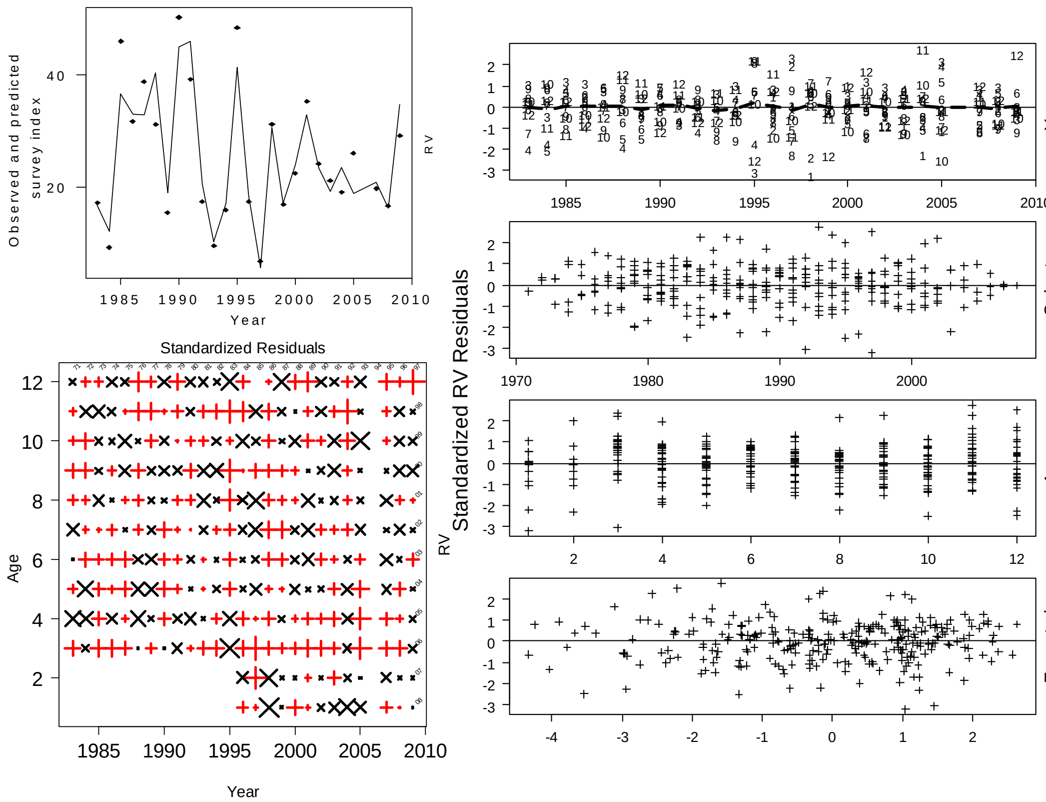

Figure 9.1.8. Top left panel: Sum of observed (points) and predicted (lines) survey indices each year. Predicted indices have been corrected for the log transformation bias. Bottom left panel: Matrix plot of residuals. Red +'s are positive and black ×'s are negative. The sizes of plotting symbols are proportional to the absolute value of the residuals. Cohorts are listed in small font along some margins. Right panels: Standard residuals vs. year, age, cohort, and predicted value. The dashed line in the top right panel indicates the average residual each year.

idea is that fishery selectivity is a combination of "domed" vs. "flat-topped" selection patterns, and the contribution of each base function changes from year to year. The selectivity model is

$$\log(s_{y,a}) = p_y \log(s_{1a}) + (1 - p_y)\log(s_{2a}), \quad 9.3.4$$

where s_{1a} and s_{2a} are the two base selection functions. The p_y weights were a parametric function of year; we assumed a logistic function,

$$p_y \equiv p(y) = \frac{\exp\{\theta(y-1994)/Y\}}{1+\exp\{\theta(y-1994)/Y\}}, Y \equiv No.\,years. \qquad 9.3.5$$

One selection function was modelled the same as in Model 1. The second function was modelled so that the log difference between it and the first function was a simple random walk, with variance equal to the sum of the variances for each selection function. The rationale for this somewhat unorthodox approach was to remove the indeterminancy in s_{1a} and s_{2a} in equation 9.3.4. Similar values for $s_{y,a}$ could be obtained with very different values for s_{1a} and s_{2a}. Our solution to this problem was to specify that $|log(s_{1a}/\ s_{2a})|$ should be not be too large, which we achieved by assuming X_a = $log(s_{1a}/\ s_{2a})$ followed a simple random walk, $X_a = X_{a-1} + \delta_a$, where δ_a are independent $N(0,Var\{log(s_{1a})\} + log(s_{2a})\})$ deviates.

Annual empirical Bayes predictions of fishery selectivity (Figure 9.1.9) suggest a switch to a more domed pattern in which ages 4 to 8 are more selected by the fishery since the moratorium in 1994–1996. This is consistent with our understanding of the selectivity of the gillnet fishery in this region.

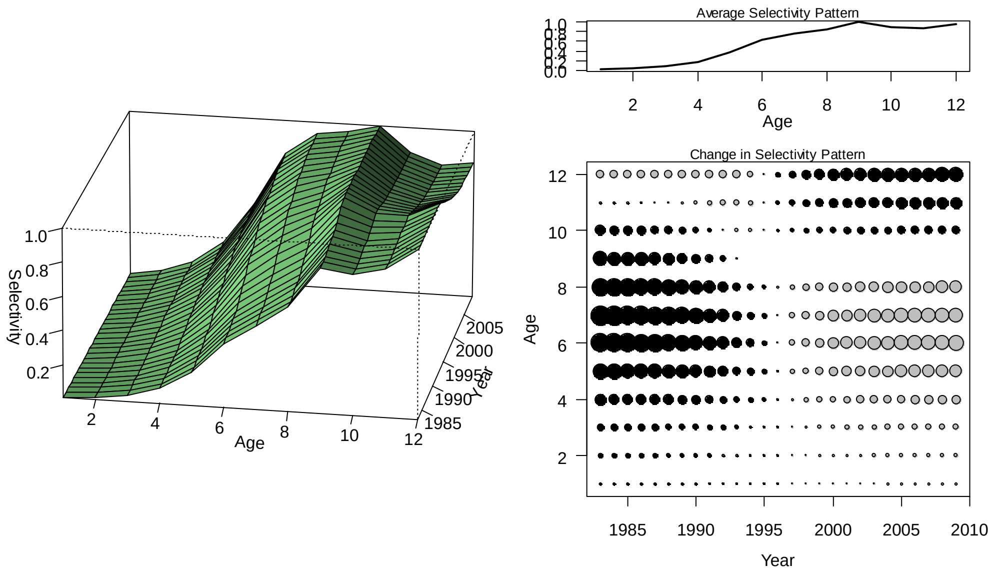

Figure 9.1.9. Left panel: Three dimensional graph of fishery selectivity, scaled to a maximum of one each year. Top right panel: Average selectivity at age over all years. Bottom right panel: Selectivity deviations from the average. Black indicates negative (i.e. less than average) and grey indicates positive.

The time-varying selectivity model did not result in substantially difference estimates of stock size or fishing mortality rates (Figure 9.1.10), but did tend to result in wider confidences for biomass and SSB except during 1983–1987. Confidence intervals for

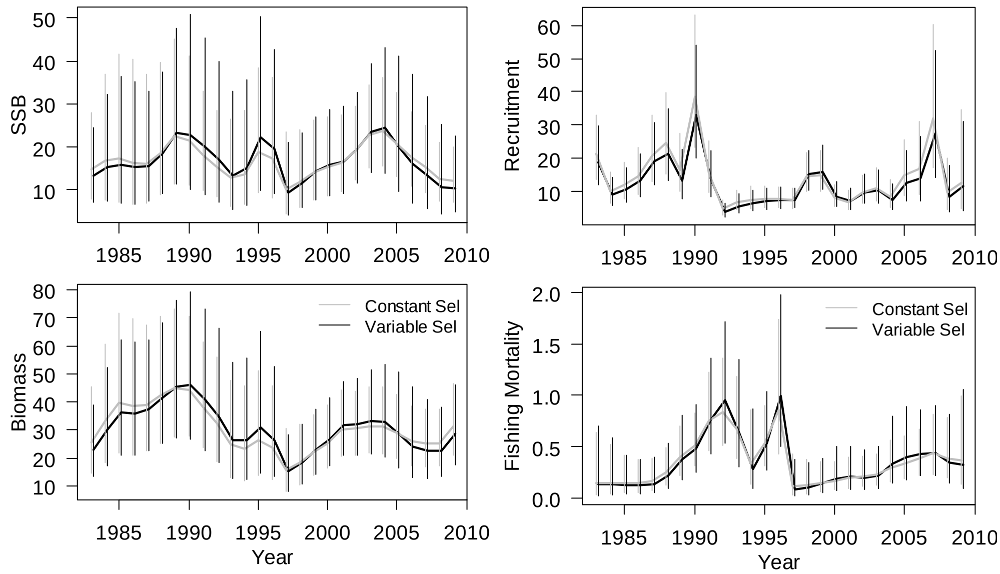

Figure 9.1.10. Left panels: Comparisons of SSB (top) and biomass (bottom) from the time-varying selectivity model (model 2; black lines) and model 1 (grey lines) in which selectivity was constant over time. Right panels: Comparisons of recruitment and average fishing mortality-at-ages 4–12. Vertical lines indicate 95% confidence intervals.

recruitment were usually similar for the two models, but confidence intervals for average fishing mortality were almost always wider for model 2. Stock size relative to 1994 levels, the limit reference point for 3Ps cod, were more uncertain especially for SSB. The additional flexibility in selectivity creates additional uncertainty about recent SSB. Increasing the amount of "dome" in the last few years can result in much higher SSB's with little change in fit to the data which is why the upper confidence interval endpoint is much higher for model 2 compared to model 1.

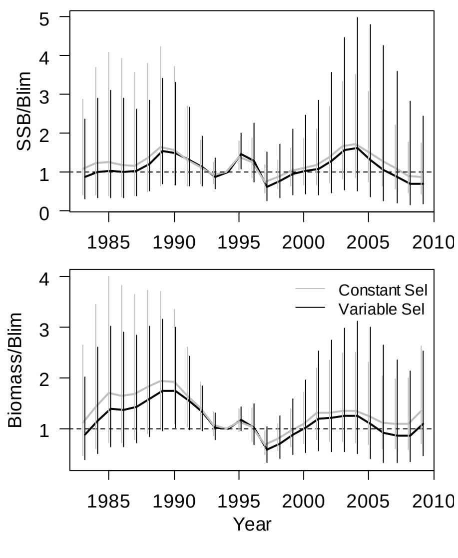

Figure 9.1.11. Comparisons of SSB (top right) and biomass (bottom panel) relative to 1994 values, which for SSB is the limit reference point for this stock. Black lines indicate results from the time-varying selectivity model (model 2) and grey lines are for model 1 in which selectivity was constant over time. Vertical lines indicate 95% confidence intervals. Dashed references lines at one are shown.

Model 2 did not fit the data much better. The improvement in fit (2×log-likelihood) was very small (0.23) which suggests that the estimated change in selectivity, although realistic in appearance, is not significant. Residuals were very similar for these two models.

9.1.5 Model 3: Random Recruitment

It is common to use random effect approaches to model recruitment and reduce the number of parameters. Often this is done using a stock recruitment model. However, there has been little evidence of a stock–recruit relationship for 3Ps cod. We decided to model log-recruitment as a normal random variable with a constant mean. Estimates of the size of recent cohorts are usually more uncertain in cohort models, and the random-effects approach we use will have the effect of "shrinking" estimates of recent recruitment to the time-series average unless the indices for recent recruitment strongly indicate otherwise. As usual, the strength of the evidence is determined via the improvement in fit to the data vs. the cost of having $Var(recruitment) = \sigma_r^2 \geq 0$ in the likelihood. Otherwise Model 3 is identical with Model 2.

Some shrinkage in recruitment predictions occurred, particularly for the 2006 year-class (Figure 9.1.12). Confidence intervals for biomass and SSB were substantially narrower in some years (e.g. 1979–1985) and never much larger. Confidence intervals for average F were very similar for the two models.

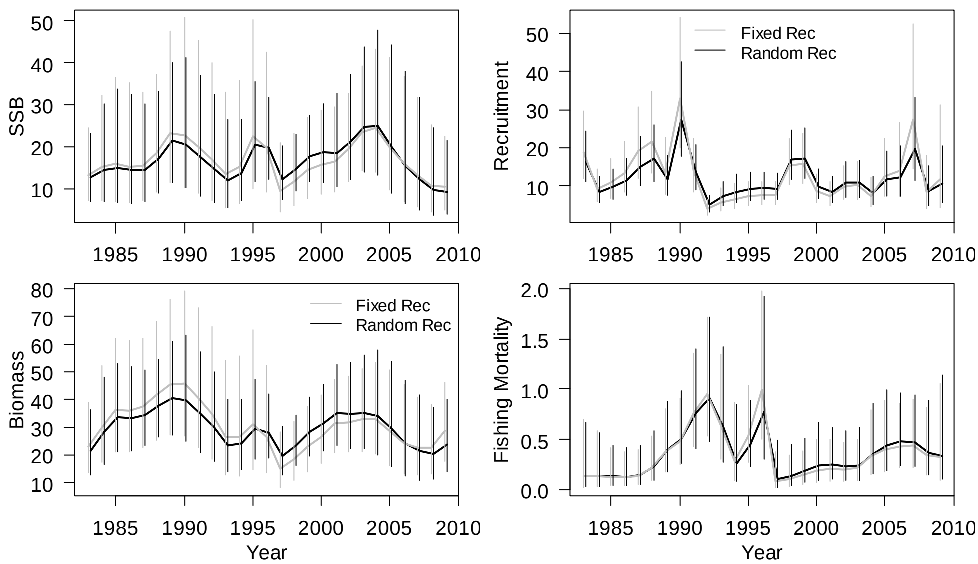

Figure 9.1.12. Left panels: Comparisons of SSB (top) and biomass (bottom) from the time-varying selectivity model (model 2; grey lines) and the random recruitment (model 3; black lines). Right panels: Comparisons of recruitment and average fishing mortality-at-ages 4–12. Vertical lines indicate 95% confidence intervals.

However, treating recruitment as a random effect resulted in wider confidence intervals in stock size relative to 1994 levels, especially during 1997–2005 (Figure 9.1.13). The biomass confidence intervals prior to 1993 were shorter for the random recruitment model. It is not obvious why these changes occur.

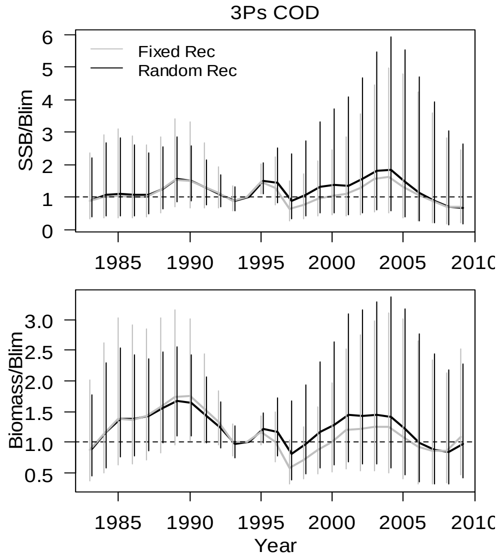

Figure 9.1.13. Comparisons of SSB (top right) and biomass (bottom panel) relative to 1994 values, which for SSB is the limit reference point for this stock. from Black lines indicate results from the random recruitment model 3 and grey lines are for time-varying selectivity model 2 in which annual recruitment was estimated as fixed parameters. Vertical lines indicate 95% confidence intervals. Dashed references lines at one are shown.

9.1.6 Recommendations

- Simulation test methods.
- Investigate if the year effects in log survey catchability can cause bias in estimated trends in stock size.
- Investigate better approaches to estimate time-varying selectivity. The procedure used here is *ad hoc*. We found that estimates and confidence intervals for trends in SSB were fairly sensitive to how the time-varying selectivity model was implemented.
- Provide stochastic stock projections. Add process error in the Z model. This is important for stochastic projections.
- Specifically for 3Ps cod, explore better ways to model fishing mortality during the short fishing moratorium in 1994–1996. It seems difficult to untangle year effects from reductions in fishing mortality that must have occurred.

9.2 Developments in SURBA-R

In addition to the developments described in Section 9.1 above, there is a requirement to develop a new implementation of the existing and widely used SURBA 3.0 code (Needle 2008, Mesnil 2009). This code was written in Fortran-90 and uses NAG library minimization routines to fit parameters. Unfortunately NAG do not permit free distribution of code which includes their routines, and so (in theory) only those with NAG libraries installed on their computers are allowed to run programs such as SURBA 3.0. While this restriction has seldom been enforced, it remains as a legal impediment which ICES should not ignore. There is thus a strong driver to redevelop SURBA for a different platform which is not subject to usage restrictions.

During the WGMG meeting, work began on a new version (SURBA-R) which is being written using the statistical package R (R Development Core Team 2005). This has several advantages: R is freely available, it avoids NAG library limitations, and the resulting function library will be useable but also modifiable. It is also much easier to generate uncertainty estimates through the R code than it was using the Fortran code. The main disadvantage of R code is that it can (in general) be quite slow to run.

Code development is as yet at an early stage, but good progress has been made. Parameter fitting is achieved using the `nls.lm` function from the `minpack.lm` library. This has proven to be faster than the `optim` function that is bundled with the R distribution. Further time savings were achieved by full vectorisation of the minimization function, as well as the reworking of all data frames as matrices, and a single case-study run now takes around 8 seconds (which compares very well with around 5 seconds for SURBA 3.0). Further work for this implementation will focus in the first instance on full replication of the SURBA 3.0 options and outputs. It is hoped that developments in SURBA-R and SURBA+ will prove to be complementary, and that a single implementation will emerge in the fullness of time.

9.3 SURBA 2.1

During the comparisons of WG assessment model use carried out for Section 5.3, we noticed that SURBA 2.1 was still being used by AFWG as the main assessment model for Norwegian coastal cod. WGMG points out that this was a trial version of SURBA which was not very well tested, and which contains unjustifiable features such as allowing catchabilities to be freely estimable parameters (which they should not be in this implementation). Use of this code as the basis for management advice is not advisable.

Recommendation: AFWG should not use SURBA 2.1 as the assessment model for Norwegian coastal cod.

10 Conclusions

10.1 New chair proposal

WGMG proposes that José De Oliveira (England) take over as the new Chair from 2010 onwards.

10.2 Future directions for WGMG

Existing approach

The approach taken for the 2009 WGMG meeting was different from previous years. Following a suggestion from Mike Sissenwine (ACOM Chair), ToRs were developed intersessionally by the WGMG Chair in consultation with benchmark and assessment Working Group chairs. The intention behind this was to ensure the relevance of WGMG output to the direct requirements of forthcoming benchmark and assessment meetings.

However, WGMG concludes that this approach has not been successful. There was considerably less buy-in to the process from benchmark and assessment Chairs than had been anticipated, with requests submitted by only five individuals – three of whom are WGMG members. The idea of finalizing the ToRs during summer has also been counter-productive. Many institutes plan their travel budget at the start of the year, and it is clear (from the low participation in this year's meeting) that a WGMG with no defined ToRs was a low priority for many who may otherwise have attended.

During this year's WGMG meeting a request was received from the ICES Secretariat to continue the current practice, and indeed to extend it so that WGMG becomes even more entwined with the benchmark process. WGMG does not agree that this is a useful way forward, for the reasons given above – it is also likely that many WGMG members will attend the 2010 benchmark meetings in other capacities in any case. We suggest a Workshop for 2010 instead: some ideas for this are outlined below.

In the longer term, it is the view of WGMG members that the group should function in an ongoing research capacity with longer term goals than trying to find quick fixes to today's assessment problems (the benchmark meetings are the place to do that, if needed). An example from the last two meetings has been developments in the SURBA survey-based assessment method: these developments have not yet reached the stage where they can be applied in assessment Working Groups, but there has been measurable and valuable progress within WGMG. There is no other forum in ICES that permits this longer term view and it would be unfortunate if it were lost.

Workshop proposal

ToRs b) and c) could not be addressed during the 2009 meeting of WGMG. Much broader participation would be required to address these ToRs. WGMG proposes that a Workshop be held in 2010 to specifically address these ToRs. The title of the workshop could be "Recent advances in stock assessment models worldwide". There has been considerable correspondence between ICES scientists on the need for such a workshop, and WGMG would seem to be the natural group to steer and convene it.

A basic task of the workshop would be to provide a reasonably comprehensive overview of "state-of-the-art" stock assessment models, both from ICES areas and elsewhere. This would include a rigorous and consistent description of the assumptions, data requirements, and estimation procedures for each method. Another task of the workshop would be to summarize the advantages and disadvantages of the various methods, and the types of stocks the methods are better suited for. This will be a difficult task that should be assisted by asking the model experts to demonstrate the improvements their models can give. That is, what problems have they fixed? WGMG suggests that model experts present case studies at the workshop that illustrate improvements, and also some simulations results that demonstrate their models provide reasonably reliable estimates and measures of uncertainty (e.g. confidence intervals) for the situations that the models were developed for. This is a first step in demonstrating the reliability of a stock assessment model. The second, and more difficult, step is to demonstrate that the model is reliable for reasonable violations of the model assumptions (i.e. is the model sufficiently robust for use by others?). This could be achieved through sensitivity analyses conducted at the workshop.

To keep the Workshop tractable, WGMG suggests that the scope of models covered be limited to single-species, age-based (or length-age-based) assessment models.

Proposed ToRs for the Workshop, and by extension, for WGMG in 2010, are given in Annex 2. A Chair for the Workshop would need to be identified.

An interim list of potential models and contact people is given below. This has a number of blank spaces, and would need revised should the proposed Workshop go ahead.

MODEL	POTENTIAL CONTACT PERSON
ADAPT	Stratis Gavaris
ADCAM	[Iceland]
A-SCALA	Mark Maunder (IATTC, La Jolla)
ASPIC	Michael Prager (Southeast Fisheries Science Center, NMFS)
B-ADAPT	Chris Darby (CEFAS, Lowestoft)
Bayesian catch-at-age model	Carmen Fernández (IEO, Vigo)
CASAL	Alastair Dunn (NIWA, New Zealand)
CLS	
Coleraine	Ray Hilborn
CSA	Benoit Mesnil (IFREMER, Nantes)
Gadget	Daniel Howell (IMR, Bergen)
ICA	Ken Patterson (EC, Brussels)
Multifan-CL	David Fournier (Otter Research)
SAD	José de Oliveira (CEFAS, Lowestoft)
SAM	Anders Nielsen (DTU-Aqua, Copenhagen)
SCCA	
SMS	Morten Vinther (DTU-Aqua, Copenhagen)
SS III	Rick Methot
SURBA	Coby Needle (MSS, Aberdeen) / Noel Cadigan (DFO, St John's)
SXSA	Dankert Skagen (IMR, Bergen)
TISVPA	Vasilyev (Murmansk)
TSA	Rob Fryer (MSS, Aberdeen)
XSA	Chris Darby (CEFAS, Lowestoft)

11 References

Allain, G., Petitgas, P., and P. Lazure. 2001. The influence of mesoscale ocean processes on anchovy (Engraulis encrasicolus) recruitment in the Bay of Biscay, estimated with a three-dimensional hydrodynamic model. *Fisheries Oceanography* 10, 151–163.

Barange, M. (Ed.). 2001. Report of the 1st Meeting of the SPACC/IOC Study Group on 'Use of Environmental Indices in the Management of Pelagic Fish Populations', September 2001, Cape Town, South Africa. GLOBEC Spec. Contrib. 5: 122 pp.

Barange, M. (Ed.). 2003. Report of the 2nd meeting of the SPACC/IOC Study Group on 'Use of Environmental Indices in the Management of Pelagic Fish Populations', December 2002, Paris, France. GLOBEC Spec. Contrib. 6: 156 pp.

Barange, M., Bernal, M., Cercole, M.C., Cubillos, L.A., Daskalov, G.M., de Moor (formerly Cunningham), C.L., De Oliveira, J.A.A., Dickey-Collas, M., Gaughan, D.J., Hill, K., Jacobson, L.D., Köster, F.W., Massé, J., Ñiquen, M., Nishida, H., Oozeki, Y., Palomera, I., Saccardo, S.A., Santojanni, A., Serra, R., Somarakis, S., Stratoudakis, Y., Uriarte, A., van der Lingen, C.D., and A. Yatsu. 2009. Current trends in the assessment and management of stocks. *In* Climate Change and Small Pelagic Fish, Chapter 9, pp 191–255. D. Checkley, J. Alheit, Y. Oozeki and C. Roy (Eds). Cambridge University Press. xii+ 372 pp.

Basson, M. 1999. The importance of environmental factors in the design of management procedures. ICES Journal of Marine Science, 56: 933–942.

Begley, J., and Howell, D. 2004. An overview of Gadget, the Globally applicable Area Disaggregated General Ecosystem Toolbox. ICES CM 2004/FF:13pp.

Borja, A., Uriarte, A., Egaña, J., Motos, L., and V. Valencia. 1998. Relationships between anchovy (Engraulis encrasicolus) recruitment and environment in the Bay of Biscay (1967–1996). *Fisheries Oceanography* 7: 375–380.

Breen, P. A., Kim, S. W., and Andrew, N. L. 2003. A length-based Bayesian stock assessment model for the New Zealand abalone Haliotis iris. *Marine Freshwater Research*, 54: 619–634.

Chen, Y., Hunter, M., Vadas, R., and Beal, B. 2003. Developing a growth-transition matrix for the stock assessment of the green sea urchin (*Strongylocentrotus droebachiensis*) off Maine. *Fishery Bulletin*, 101: 737–744.

Cochrane, K. L., and A. M. Starfield, 1992. The potential use of predictions of recruitment success in the management of the South African anchovy resource. *In* Benguela Trophic Functioning. Payne, A. I. L., Brink, K. H., Mann, K. H. and R. Hilborn (Eds). *South African Journal of Marine Science*, 12: 891–902.

Cruywagen, G. 1997. The use of generalised linear modelling to determine interannual and inter-area variation of growth rates: the cape rock lobster as example. *Fisheries Research*, 29: 119–131.

Darby, C. D., and Flatman, S. 1994. Virtual Population Analysis: version 3.1 (Windows/DOS) user guide, CEFAS, Lowestoft, UK. Information Technology Series No 1.

De Leo, G. A., and Gatto, M. 1995. A size and age-structured model of the European eel (*Anguilla anguilla* L.). *Canadian Journal of Fisheries and Aquatic Sciences*, 52:1351–1367.

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

De Oliveira, J. A. A., Uriarte, A., and B. A. Roel. 2005. Potential improvements in the management of Bay of Biscay anchovy by incorporating environmental indices as recruitment predictors. *Fisheries Research*, 75: 2–14.

de Valpine, P. and Hilborn, R. 2005. State-space likelihoods for nonlinear fisheries time-series. *Canadian Journal of Fisheries and Aquatic Sciences*, 62: 1937–1952.

DeLong, A. K., Collie, J. S., Meise, C. J., and Powell, J. C. 2001. Estimating growth and mortality of juvenile winter flounder, *Pseudopleuronectes americanus*, with a length-based model. *Canadian Journal of Fisheries and Aquatic Sciences*, 58: 2233–2246.

Deriso, R. B., Barnes, J. T., Jacobson, L. D., and P. R. Arenas. 1996. Catch-at-age analysis for Pacific sardine (*Sardinops sagax*), 1983–1995. *Rep. Calif. coop. oceanic Fish. Invest*. 37: 175–187.

DFO. 2009. Stock Assessment of Subdivision 3Ps cod. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2009/008.

Drouineau, H. 2008. *Développement and ajustement d'un modèle de dynamique des populations structuré en longueur and spatialisé appliqué au stock nord de merlu (Merluccius merluccius).* PhD thesis, Rennes: Agrocampus.

Eveson, J. P., Laslett, G. M., and Polacheck, T. 2004. An integrated model for growth incorporating tag-recapture, length frequency, and direct aging data. *Canadian Journal of Fisheries and Aquatic Sciences*, 61: 292–306.

FLR Team. 2006. FLR: Fisheries modelling in R. Version 1.2.1. Initial design by L.T. Kell, P. Grosjean, http://www.flr-project.org/doku.php.

Fournier, D. A., Hampton, J., and Sibert, J. R. 1998. MULTIFAN-CL: a length-based, agestructured model for fisheries stock-assessment, with application to South Pacific albacore, *Thunnus alalunga*. *Canadian Journal of Fisheries and Aquatic Sciences*, 55: 2105–2116.

Frøysa, K. G., Bogstad, B., and Skagen, D. W. 2002. Fleksibest: an age-length structured fish stock assessment model. *Fisheries Research*, 55: 87–101.

Fryer, R. 2002. TSA: is it the way? Appendix D in the Report of the ICES Working Group on Methods of Fish Stock Assessment, December 2001. ICES CM 2002/D:01, pp. 86–93.

Fu, C., and T. J. Quin II. 2000. Estimability of natural mortality and other population parameters in a length-based model: *Pandalus borealis* in Kachemak Bay, Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, 57: 2420–2432.

Gudmundsson, G. 1987. Time series models of fishing mortality rates. ICES C.M. d:6.

Gudmundsson, G. 1994. Time series analysis of catch-at-age observations. *Appl. Statist*. 43:117–126.

Hilborn, R. and C. J. Walters. 1992. *Quantitative fisheries stock assessment: choice dynamics and uncertainty*. Chapman and Hall, New York: [xv] + 570pp.

ICES-ACFM 1995. Guidance to Assessment Working Groups from ACFM and the Secretariat: The Blue Pages. ICES, March 1995.

ICES-SGASAM. 2006. Report of the Study Group on Age-Length Structured Assessment Models (SGASAM). ICES CM 2006/RMC: 10. 73pp.

ICES-SGGROMAT. 2002. Report of the Study Group on Growth, Maturity and Condition in Stock Projections. ICES Headquarters, 5–10 December 2002. ICES CM 2003/D:01.

ICES-SGGROMAT. 2004. Report of the Study Group on Growth, Maturity and Condition in Stock Projections. University of Aberdeen, Scotland, 19–23 January 2004. ICES CM 2004/XXXX.

ICES-SGMAS. 2008. Report of the Study Group on Management Strategies (SGMAS). 17–21 November 2008, Lisbon, Portugal. ICES CM 2008/ACOM:24: 74 pp.

ICES-WGDEEP. 2009. Report of the Working Group on the Biology and Assessment of Deep Sea Fisheries Resources (WGDEEP), 9–16 March 2009, Copenhagen, Denmark. ICES CM 2009/ACOM:14. 511pp.

ICES-WGEF. 2006. Report of the Working Group on Elasmobranch Fishes (WGEF), 14–21 June 2006, ICES Headquarters. ICES CM 2006/ACFM:31.

ICES-WGHMSA. 2000. Section 10: Anchovy - Sub-area VIII. In: Report of the working group on the assessment of mackerel, horse mackerel, sardine and anchovy (WGMHSA). 14–23 September 1999, ICES Headquarters, Copenhagen. ICES CM 2000/ACFM:5: 409–466.

ICES-WGMG. 1985. Report of the Working Group on Methods of Fish Stock Assessments. Copenhagen, 11–15 June 1984. ICES Cooperative Research Report, 133: 56.

ICES-WGMG. 1991. Report of the Working Group on Methods of Fish Stock Assessments, St. John's (Canada), 20–27 June 1991, International Council for the Exploration of the Sea, C.M. 1991/Assess:25.

ICES-WGMG. 1993a. Reports of the Working Group on Methods of Fish Stock Assessments. Copenhagen, 9–16 June 1987; Reykjavik, 6–12 July 1988; Nantes, 10–17 November 1989. ICES Cooperative Research Report, 191: 249.

ICES-WGMG. 1993b. Report of the Working Group on Methods of Fish Stock Assessment. Copenhagen, 3–10 February 1993, International Council for the Exploration of the Sea, C.M. 1993/Assess:12.

ICES-WGMG. 2008. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 7–16 October 2008, Woods Hole, USA. ICES CM 2008/RMC:03. 147 pp.

ICES-WGMHSA. 2001. Section 11: Anchovy - Sub-area VIII. In: Report of the working group on the assessment of mackerel, horse mackerel, sardine and anchovy (WGMHSA). 14–23 September 2000, ICES Headquarters, Copenhagen. ICES CM 2001/ACFM:06: 340–414.

ICES-WGMHSA. 2002. Section 11: Anchovy - Sub-area VIII. In: Report of the working group on the assessment of mackerel, horse mackerel, sardine and anchovy (WGMHSA). 4–13 September 2001, ICES Headquarters, Copenhagen. ICES CM 2002/ACFM:06: 360–421.

ICES-WGNSSK. 2006. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2006/ACFM:35.

ICES-WGNSSK. 2007. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2007/ACFM:18 and 30.

ICES-WGNSSK. 2008. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2008/ACFM:09.

ICES-WGNSSK. 2009. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2009/ACFM:10.

ICES-WKARRG. 2007. Report of the Workshop on Age Reading of Roundnose Grenadier (WKARRG). Boulogne sur Mer, 4–7 September 2007. ICES CM 2007/ACE:36. 50pp.

ICES-WKEFA. 2007. Report of the Workshop on the Integration of Environmental Information into Fisheries Management Strategies and Advice (WKEFA). 18–22 June 2007, ICES Headquarters, Copenhagen. ICES CM 2007/ACFM:25: 182 pp.

Jacobson, L. D., and A. D. MacCall. 1995. Stock-recruitment models for Pacific sardine (*Sardinops sagax*). *Canadian Journal of Fisheries and Aquatic Sciences*, 52: 566–577.

Jacobson, L. D., Bograd, S. J., Parrish, R. H., Mendelssohn, R., and F. B. Schwing. 2005. An ecosystem-based hypothesis for climatic effects on surplus production in California sardine (*Sardinops sagax*) and environmentally dependent surplus production models. *Canadian Journal of Fisheries and Aquatic Sciences*, 62: 1782–1796.

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

Kristensen, K., Lewy, P., and Beyer, J. 2006. How to validate a length-based model of single species fish stock dynamics. *Canadian Journal of Fisheries and Aquatic Sciences*, 63:2531–2542.

Laslett, G. M., Eveson, J. P., and Polacheck, T. 2002. A flexible maximum likelihood approach for fitting growth curves to tag-recapture data. *Canadian Journal of Fisheries and Aquatic Sciences*, 59: 976–986.

MacKenzie, B. R. M. and F. W. Köster. 2004. Fish production and climate: sprat in the Baltic Sea. *Ecology* 85: 784–794.

MacKenzie, B. R. M., Horbowy, J., and F. W. Köster. 2008. Incorporating environmental variability in stock assessment: predicting recruitment, spawner biomass, and landings of sprat (Sprattus sprattus) in the Baltic Sea. *Canadian Journal of Fisheries and Aquatic Sc-ences*, 65: 1334–1341.

Maunder, M. N., and Watters, G. M. 2003. A general framework for integrating environmental time series into stock assessment models: model description, simulation testing, and example. *Fishery Bulletin*, 101: 89–99.

Maunder, M. N., and Watters, G. M. 2003. A-SCALA: un analisis estadistico de captura a talla estructurado por edad para la evaluacion de las poblaciones de atunes en el oceano pacifico oriental. *Bulletin Inter-American Tropical Tuna Commission*, 22: 531–581.

Megrey, B. A., Lee, Y.-W., and S. A. Macklin. 2005. Comparative analysis of statistical tools to indentify recruitment-environment relationships and forecast recruitment strength. *ICES Journal of Marine Science*, 62: 1256–1269.

Mesnil, B., Cotter, A. J. R., Fryer, R. J., Needle, C. L. and Trenkel, V. M. 2009. A review of fishery independent assessment models, and initial evaluation based on simulated data, *Aquatic Living Resources* 22: 207–216.

Myers, R. A. 1998. When do environment-recruitment correlations work? *Revs. Fish Biol. Fish*. 8: 285–305.

Needle, C. L. 2002. Recruitment models: diagnosis and prognosis. *Reviews in Fish Biology and Fisheries*, 11: 95–111.

Needle, C. L. 2008. Survey-based fish stock assessment with SURBA. Course given at the North-West Atlantic Fisheries Centre (DFO), St. John's, Newfoundland, Canada, 3–4 September 2008.

NRC. 1998. *Improving Fish Stock Assessments*. Washington D.C., National Academy Press.

Panfili, J., Pontual, H. (de), Troadec, H., Wright, P.J. 2002. Manual of fish schlerochronology. Coédition Ifremer-IRD, 464p.

Patterson, K. R., and Melvin, G. D. 1996. Integrated Catch-At-Age Analysis, Version 1.2, Scottish Fisheries Research Report, 58.

Pawlowski, L., and Lorance, P. 2009. Effect of discards on roundnose grenadier stock assessment in the Northeast Atlantic. *Aquatic Living Resources*, DOI: 10.1051/alr/2009040.

PFMC. 1998. Amendment 8 to the Northern Anchovy Fishery Management Plan. Pacific Fishery Management Council, 2130 SW Fifth Ave, Suite 224, Portland, OR 97201. [Accessed online in October 2009: http://www.pcouncil.org/cps/cpsfmp.html]

Pilling, G. M., Kirkwood, G. P., and Walker, S. G. 2002. An improved method for estimating individual growth variability in fish, and the correlation between von Bertalanffy growth parameters. *Canadian Journal of Fisheries and Aquatic Sciences*, 59: 424–432.

Punt A. E. 2008. Refocusing stock assessment in support of policy evaluation. In: Tsukamoto K, Kawamura T, Takeuchi T, Beard TD, Jr, Kaiser MJ (eds). Fisheries for Global Welfare and Environment, 5th World Fisheries Congress 2008. TERRAPUB, Tokyo: 139–152.

Punt, A. E. 2003. The performance of a size-structured stock assessment method in face of 151 spatial heterogeneity in growth. *Fisheries Research*, 65: 391–409.

Punt, A. E., and Walker, T. I. 1998. Stock assessment and risk analysis for the school shark (*Galeorhinus galeus*) off southern Australia. *Marine and Freshwater Research* 48, 719–31.

R Development Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Rosenberg, A. A., Kirkwood, G. P., Cook, R. M., and Myers, R. A. 1992. Combining information from commercial catches and research surveys to estimate recruitment: A comparison of methods. *ICES Journal of Marine Science*, 49: 379–387.

Sainsbury, K. J. 1980. Effect of individual variability on the von Bertalanffy growth equation. *Canadian Journal of Fisheries and Aquatic Sciences*, 377: 241–247.

Schirripa, M. J. 2007. Status of the Sablefish Resource off the Continental U.S. Pacific Coasts in 2007. Agenda Item E.6.a, Supplemental Revised Attachment 1, June 2007. [Accessed online in October 2009: http://www.pcouncil.org/bb/2007/0607/E6a_REV_ATT1.pdf]

Shepherd, J. G. 1999. Extended survivors analysis: An improved method for the analysis of catch-at-age data and abundance indices. *ICES Journal of Marine Science*, 56: 584–591.

Smith, B. D., and Botsford, L. W. 1998. Interpretation of growth, mortality, and recruitment patterns in size-at-age, growth increment, and size frequency data. Proceedings of the North Pacific Symposium on Invertebrate Stock Assessment and Management in *Can. J. Fish. Aquat. Sci*., 125: 125–139.

Smith, B. D., Botsford, L. W., and Wing, S. R. 1998. Estimation of growth and mortality parameters from size frequency distributions lacking age patterns: the red sea urchin (*Strongylocentrotus franciscanus*) as an example. *Canadian Journal of Fisheries and Aquatic Sciences*, 55: 1236–1247.

Sullivan, P. J., Lai, H., and Galluci, V. F. 1990. A catch-at-length analysis that incorporates a stochastic model of growth. *Canadian Journal of Fisheries and Aquatic Sciences*, 47:184–198.

Uriarte, A., Roel, B.A., Borja, A., Allain, G., and C. M. O'Brien. 2002. Role of Environmental indices in determining the recruitment of the Bay of Biscay anchovy. ICES CM 2002/O:25.

Walters, C. J. 1989. Value of short-term forecasts of recruitment variation for harvest management. *Canadian Journal of Fisheries and Aquatic Sciences*, 46: 1969–1976.

Annex 1: List of participants

Name	Address	Phone/Fax	Email
Anders Nielsen	DTU-Aqua Charlottenlund Slot Jægersborg Allé 1 DK-2920 Charlottenlund Denmark		an@aqua.dtu.dk
Benoît Mesnil	IFREMER BP 21105 F44311 – Nantes Cedex 3 France	Tel: +33 240 374 009	Benoit.Mesnil@ifremer.fr
Coby Needle *(Chair)*	MSS Marine Laboratory PO Box 101 375 Victoria Road Aberdeen AB11 9DB Scotland	Tel: +44 1224 295456 Fax: +44 1224 295511	needlec@marlab.ac.uk
José De Oliveira	CEFAS Lowestoft Laboratory Pakefield Road Lowestoft Suffolk NR33 7BA UK	Tel: +44 1502 527727 Fax: +44 1502 524511	jose.deoliveira@cefas.co.uk
Lionel Pawlowski	IFREMER 8 Rue François Toullec 56100 Lorient France	Tel: +33 297 873846 Fax: +33 297 873836	Lionel.Pawlowski@ifremer.fr
Noel Cadigan	DFO Northwest Atlantic Fisheries Science Center 80 White Hills Road East PO Box 5667 St. John's, NL Canada	Tel: +1 709 772 5028	Noel.Cadigan@dfo-mpo.gc.ca

Verena Trenkel (IFREMER, Nantes) also attended the meeting during one morning in the first week, to participate in the discussion on spurdog assessment methods.

The full list of WGMG members is given below for comparison:

Alberto Murta	Dankert Skagen	Jose de Oliveira	Noel Cadigan
Anders Nielsen	David Orr	Knut Korsbrekke	Paul Rago
Benoit Mesnil	Dmitri Vasilyev	Laurence Kell	Peter Lewy
Carl O'Brien	Dorleta Garcia	Leire Ibaibarriaga	Sam Subbey
Carmen Fernandez	Fatima Cardador	Lionel Pawlowski	Sarah B. M. Kraak
Chris Darby	Frans van Beek	Liz Brooks	Sigurdur Thor Jónsson
Christopher Legault	Geert Aarts	Manuela Azevedo	Stéphanie Mahevas
Claire Welling	Jan Arge Jacobsen	Mario Rui Pinho	Tim Miller
Coby Needle	Jan Jaap Poos	Mark Terceiro	Tobias van Kooten
Colin Millar	Jim Kristmanson	Martin Aranda	Yuri A. Kovalev
Colm Lordan	Joachim Gröger	Michel Bertignac	Yuri Efimov
Daniel Howell	Jon Brodziak	Mike Sissenwine	

Annex 2: WGMG Terms of Reference for the next meeting

The **ICES Workshop on Reviews of Recent Advances in Stock Assessment Models Worldwide "Around The World In AD Models"** (WKADSAM) co-chaired by Coby Needle, UK*, and Chris Legault*, USA) will meet in Nantes, France, for 5 days in September 2010 to collate, review and comment on stock assessment methods currently in use around the world.

This will be part of the ICES initiative on stock assessments methods.

The workshop will:

a) Determine the key techniques and approaches used to assess fish stocks
b) Consider inter alia utility, ease of use, estimation procedures, robustness, suitability to different data richness, applicability to data poor situations, and relevance of assumptions in the models
c) Summarize the advantages and disadvantages of the various methods, and describe the appropriate use.
d) Comment on demonstrations by model developers of the utility of methods with case studies and simulated data sets, focussing on the question: What problem has the method fixed?
e) Prepare the groundwork for a following workshop in 2011 or 2012 (see initiative plan below)

Supporting Information

Priority	The work of this group is essential to ICES to progress in the development of methods for fish stock assessment and advice.
Scientific justification	This workshop forms the essential first part of a proposed ICES initiative to carry out a review of the assessment methods currently in use worldwide. Knowledge gained at the workshop will be used to inform plans for the second and third phases, respectively a major international conference on stock assessment methods (in the second year) and a published review (in the third year). The purpose of the initiative is develop, improve and otherwise modernise the stock assessment methods used by ICES assessment working groups, and thereby improve the provision of scientific fisheries management advice in years to come.
Resource requirements	None.
Participants	ICES does not intend to carry out this review in isolation. It has already approached and received support from FAO and is about to approach ICCAT, IATTC, CCSBT, NAFO and IPHC. It is also proposed to approach the network of Regional Fisheries Organizations Secretariats hosted by FAO. In addition individual institutes/countries will be approached: in Australia; New Zealand; Japan; South Africa plus non- ICES scientists in Russia, USA and Canada.
Secretariat facilities	None required in 2010

Financial	This is part of the initiative on stock assessment methods.
Linkages to advisory committees	The initiative will be jointly managed by ICES ACOM and SCICOM with input from the ICES Working Group on Methods of Fish Stock Assessments.
Linkages to other committees or groups	WKADSAM will report to WGMG and the SCICOM by Dec 2010.
Linkages to other organizations	Linkages to other organizations: see above.

Draft plan for ICES initiative

ICES initiative: Review of stock assessment methods used around the world

ICES, along with other partner fisheries organizations, plans to carry out a joint review of stock assessment methods over the next two to three years.

Justification

There have been many recent advances in fish stock assessment methods and techniques. Many of these advances are conceptual and others are technological. ICES has been slow to incorporate many of these developments into its advisory system. ICES used to be a world leader in the development of stock assessment methods and in some areas still is, but the majority of the ICES stock assessments now use methods that can be considered out-dated, compared to some of those used outside ICES. Most of the current methods fail to fully utilize the available data resources. In other cases, the lack of standard catch-at-age and classic fisheries independent time-series often results in no assessment being made, whilst useful information for these "data poor" stocks does exist. As the client organizations of ICES require a broader portfolio of fisheries advice, as well as integrated regional advice, ICES is finding that some of the stock assessment methods it uses are failing to provide the necessary basis for such advice.

ICES needs an approach by which it can re-invigorate the stock assessment methods it uses, and stimulate the development of new techniques and concepts. As this must be done without re-inventing the wheel, ICES requires a review of methods used around the world for fish stock assessment. It is hoped that this review will advance not just ICES knowledge but also the operation of its stock assessment experts and the advisory system as a whole.

Objective

To carry out a review of state-of-the-art stock assessment methods used around the world. The review will result in a major publication of the findings and a repository of online, free, robust and tested stock assessment methods.

Partners

ICES does not intend to carry out this review in isolation. It has already approached and received support from FAO and is about to approach ICCAT, IATTC, CCSBT, NAFO and IPHC. It is also proposed to approach the network of Regional Fisheries Organizations Secretariats hosted by FAO. In addition individual institutes/countries will be approached: in Australia; New Zealand; Japan; South Africa plus non- ICES scientists in Russia, USA and Canada.

The initiative will be jointly managed by ICES ACOM and SCICOM with input from the ICES Working Group on Methods of Fish Stock Assessments. The first workshop will be chaired by Coby Needle (UK) and Chris Legault (USA). The larger workshop will be convened by (to be determined). The final review will be edited by Dickey-Collas, Needle, and …????.

Approach

The review will take the form of two workshops leading to a publication of the review of state-of-the-art methods and an online repository of codes, manuals and working data sets.

The format and style of the publication will be determined at the first workshop. At this first workshop the key techniques and approaches will be identified and the sections of the review will be allocated to contributors (as work packages). Importantly the review will not just list the stock assessment methods, but will focus and comment on them too. The criteria for such comments will probably include utility, ease of use, estimation procedures, robustness, suitability to different data richness, applicability to data poor situations, relevance of assumptions etc. Another task of the workshop would be to summarize the advantages and disadvantages of the various methods, and describe the appropriate use. This challenging task will be assisted by asking the experts to demonstrate the utility of methods with case studies and simulated data sets. The methods will also be challenged with data sets that violate their model assumptions.

The second workshop is planned to be a major world conference on stock assessment methods with invited and contributing scientists. The objective of the conference would be to determine the state-of-the-art for stock assessment methods around the world. It is likely that each session will be method based (targeting a work package), but importantly the session will combine presentations with active workshops to determine the approaches for major chapters in the final publication.

The review will be published either as an FAO or ICES publication. It is important to aim high with regards to the quality of the contributions, the potential readership and its readability. The target group for the review will be scientists that use stock assessments. This means that the authors of the review (probably stock assessment methods developers) write in a manner that will be appropriate for the users of their products. In addition to the published review, an online repository of current techniques and methods will be set up. This will be held by ICES and those methods deposited will be as freely available possible to all. Worked examples, with appropriate data sets, will also be made available.

The publication and depository will become obsolete over time but it is hoped that the review will stimulate scientists within ICES to think beyond their current approaches and also facilitate a sharing of stock assessment methods and approaches across the world. It will place ICES back at the centre of stock assessment method development.

Products

The final product will be an FAO or ICES published review of state-of-the-art stock assessment methods and an ICES-held repository of stock assessment methods. This will be delivered in 2013.

A bi-product will be the re-ignition of the development and sharing of stock assessment methods in the ICES community and beyond. In addition, the initiative will place ICES back at the centre of stock assessment technical and conceptual development.

Resources

- ICES data centre (web support for initiative and repository) — 200 hours secretariat time
- Support for large conference (including travel for invited speakers) — €30 000
- Large conference administration — 200 hours secretariat time
- Support for publication — €10 000
- Support for marketing and awareness building — 50 hours secretariat time

Time line

The review will take place over 24 or 36 months.

Starting with the first workshop in September 2010

The second workshop (the world conference) will take place autumn 2011 or spring 2012.

The review and repository will be completed by end 2012/start 2013.

The review will be published 6 months after submission (in 2013).

The ICES Working Group on Methods of Fish Stock Assessment (WGMG) chaired by José de Oliveira*, UK will work by correspondence during 2010 to:

a) work with SCICOM and ACOM to build the ICES Initiative on Stock Assessment Methods;

b) prepare by correspondence for the forthcoming Workshop on Reviews of Recent Advances in Stock Assessment Models Worldwide in September 2010;

c) consider the structure and function of a science group on stock assessment methods within ICES.

WGMG will report by 1 December 2010 (via SSGSUE) for the attention of SCICOM and ACOM.

Supporting Information

Priority:	The work of this group is essential to ICES to progress in the development of methods for fish stock assessment and advice.
Scientific justification	The approach taken for the 2009 WGMG meeting was different from previous years. ToRs were developed intersessionally by the WGMG Chair in consultation with benchmark and assessment Working Group chairs. The intention behind this was to ensure the relevance of WGMG output to the direct requirements of forthcoming benchmark and assessment meetings. However, WGMG concluded in 2009 that this approach had not been successful. There was considerably less buy-in to the process from benchmark and assessment Chairs than had been anticipated, with requests submitted by only five individuals – three of whom are WGMG members. The idea of finalizing the ToRs during summer has also been counter-productive. Many institutes plan their travel budget at the start of the year, and it is clear (from the low participation in this year's meeting) that a WGMG with no defined ToRs was a low priority for many who may otherwise have attended. During this year's WGMG meeting a request was received from the ICES Secretariat to continue the current practice, and indeed to extend it so that WGMG becomes even more entwined with the benchmark process. WGMG does not agree that this is a useful way forward, for the reasons given above – it is also likely that many WGMG members will attend the 2010 benchmark meetings in other capacities in any case. We suggest a Workshop for 2010 instead and further involvement in development of the ICES Initiative on Stock Assessment Methods. In the longer term, it is the view of WGMG members that the group should function in an ongoing research capacity with longer term goals than trying to find quick fixes to today's assessment problems (the benchmark meetings are the place to do that, if needed). An example from the last two meetings has been developments in the SURBA survey-based assessment method: these developments have not yet reached the stage where they can be applied in assessment Working Groups, but there has been measurable and valuable progress within WGMG. There is no other forum in ICES that permits this longer-term view and it would be unfortunate if it were lost.
Resource requirements	None
Participants	Research scientists involved in stock assessment methods from the ICES area and invited scientists from outside the ICES area.

Secretariat facilities	None, other than formatting and publishing of the final report.
Financial:	No financial implications
Linkages tTo Advisory Committees	ACOM has strongly supported the work of this group and has worked actively in formulating the ToRs for recent meetings. WGMG will also report to ACOM in 2010.
Linkages to other Committees or Groups	WGMG will report to the SCICOM by December 2010.
Linkages to other Organisations	There is similar work going on within ICCAT and NAFO. Coordination should be assured.

Annex 3: Recommendations

Recommendation	For follow up by:
XSA year shrinkage: Benchmark groups should consider a proper time-series methodology to adjust or estimate the weight given to shrinkage.	Benchmark WGs.
XSA year shrinkage: Determining the amount of shrinkage by minimizing retrospective patterns is **not** recommended because it could lead to seriously biased estimates of stock size or their trends.	Benchmark WGs.
XSA year shrinkage: Whether shrinkage is "light" or "heavy" should be determined with reference to the actual scaled weights in the XSA diagnostics, not just the shrinkage SE.	Benchmark WGs.
XSA shrinkage: Although they serve a different purpose, year- and age-shrinkage are switched on in the same menu and a single CV is applied in the standard VPA suite. A version where the two are disconnected is available (C. Darby, CEFAS) and should be implemented in ICES.	ICES Secretariat.
XSA iterations: FLXSA output should include the number of iterations taken to reach the solution.	ICES Secretariat.
XSA iterations: Convergence behaviour of XSA (and indeed all models) should always be checked.	Benchmark and assessment WGs.
XSA iterations: Slow convergence and/or high sensitivity to the number of iterations could indicate that XSA is not suitable for the stock concerned. Alternative models that do not rely on *ad hoc* assumptions and algorithms should be explored for these cases.	Benchmark WGs.
ALK Uncertainty: WGMG considers age-based assessments unreliable methods for roundnose grenadier and suggests development on life-stage-structured and/or production approaches. Length-based approaches may be limited for this stock due to the lack of data on growth and uncertainties on age reading.	WGDEEP and relevant benchmark WGs.
State-space assessment model (SAM): This is a rigorously constructed model which should be considered as a valid alternative to existing methods, since it is able to quantify estimation uncertainty.	Benchmark WGs.
Survey-based assessment methods: SURBA 2.1 should not be used for the assessment of Norwegian coastal cod, as it is not well-tested or reliable.	AFWG and relevant benchmark WGs.

Annex 4: Stock simulation for testing XSA iteration convergence

The simulation used in Section 5 to test XSA iteration convergence properties is based on that developed by NRC (1998). The implementation described here is coded in Fortran-90, but versions in R and Excel are also available (Needle, pers. comm). The key points are:

- Ages: 1 to 10+. Years: 2001 to 2030.
- Natural mortality $M_{a,y}$ is based on the values used in past meetings of WGNSSK for North Sea cod, from which $M_a^{cod} = (0.8, 0.35, 0.25, 0.2, 0.2, \ldots)$. For each year of the simulation, these values are randomly modified using a year-specific multiplier $M_{a,y} = M_a^{cod} M_y^*$, with $M_y^* \sim U(1 - M_{range}, 1 + M_{range})$. In the simulation used in Section 5, $M_{range} = 0.1$.
- Mean stock weights-at-age (also used for catch weights-at-age) $W_{a,y}$ for age a and year y are derived from a cohort-based von Bertalanffy model:

$$W_{a,y} = W_\infty \left[1 - \exp\left(-a\kappa_y^W \right) \right]^{\beta_y^W} .$$

  In these simulations, $W_\infty = 5.0$, $\kappa_y^W \sim U(3.0 - \kappa_{range}, 3.0 + \kappa_{range})$, $\beta_y^W \sim U(3.0 - \beta_{range}, 3.0 + \beta_{range})$, $\kappa_{range} = 0.15$ and $\beta_{range} = 1.0$.
- The proportions mature at age $\mathrm{Mat}_{a,y}$ are similarly given by a cohort-based logistic model:

$$\mathrm{Mat}_{a,y} = \frac{1}{1 + \exp\left[-\beta_y^{\mathrm{Mat}} \left(a - \alpha_y^{\mathrm{Mat}} \right) \right]} .$$

  Here $\alpha_y^{\mathrm{Mat}} \sim U(5.0 - \alpha_{range}^{\mathrm{Mat}}, 5.0 + \alpha_{range}^{\mathrm{Mat}})$ and $\beta_y^{\mathrm{Mat}} \sim U(1.65 - \beta_{range}^{\mathrm{Mat}}, 1.65 + \beta_{range}^{\mathrm{Mat}})$. In these simulations, $\alpha_{range}^{\mathrm{Mat}} = \beta_{range}^{\mathrm{Mat}} = 0.5$.
- Fishing mortality $F_{a,y}$ is modelled as a function of selectivity $s_{a,y}$, catchability q_y and effort E_y, so that $F_{a,y} = s_{a,y} q_y E_y$. In turn, these variables are generated using:

$$s_{a,y} = \frac{1}{1 + \exp\left[-\beta_{s,y} \left(a - \alpha_{s,y} \right) \right]} \exp\left[\sigma_s \varepsilon_{y,s} - \tfrac{1}{2} \sigma_s^2 \right]$$

  where

$$\alpha_{s,y} = \begin{cases} 4.0, & y < y_s \\ 4.0 + 2.0\left[1.0 - \exp\left(-0.45 \left(y - y_s \right) \right) \right], & y \geq y_s \end{cases}$$

  and

$$\beta_{s,y} = \begin{cases} 0.6, & y < y_s \\ 0.8, & y \geq y_s \end{cases}$$

  Variance parameter σ_s is set to 0.1, and the switch year y_s is randomly chosen between 2026 and 2030 (here we are modelling a change in selectivity towards the end of the time-series). In addition, uniform random noise (maximum ± 0.1) is applied to $\alpha_{s,y}$. Catchability is determined via

$$q_y = q_0 B_0^{q_1 - 1} \exp\left[r_q \left(y - 2001 \right) + \sigma_q \varepsilon_{q,y} - \tfrac{1}{2} \sigma_q^2 \right],$$

where

$$q_0 = 0.08,$$
$$q_1 = 0.40,$$
$$B_0 = 12000,$$

and $\sigma_q = 0.2$, $r_q = 0.02$. The random element $\varepsilon_{q,y} \sim N\left(0,1\right)$. Finally effort is given by

$$E_y = E_0 E_1 \nu_E$$

where

$$E_0 = 4000,$$
$$E_1 = \begin{cases} 0.8, & y \le y_{s,E} \\ 1.2, & y > y_{s,E} \end{cases}$$
$$\nu_E \sim U\left(1 - \nu_{range,E}, 1 + \nu_{range,E}\right).$$

Here the switch year y_s is actually set to 2030 (so there is no predetermined increase in effort), and $\nu_{range,E} = 0.25$.

- Recruitment for each year is generated as a log residual to one of seven (randomly selected) underlying models (Ricker, Beverton–Holt, power, Shepherd, Saila-Lorda, changepoint and mixed; Needle 2002) from which the recruiting abundance-at-age 1 is derived. The time-series of log residuals is further constrained by a weak autoregressive assumption, and random noise is applied to provide realistic levels of variability.
- Finally, abundance and catch are produced using the standard exponential decay and Baranov equations respectively.